Partial Identification of Functionals of the Joint Distribution of "Potential Outcomes"*

Yanqin Fan
Department of Economics
University of Washington
Box 353330
Seattle, WA 98195

Emmanuel Guerre
School of Economics and Finance
Queen Mary, University of London
Mile End Road
London E1 4NS, United Kingdom

Dongming Zhu
School of Economics & Key Laboratory of Mathematical Economics
Shanghai University of Finance and Economics
777 Guoding Road, Yangpu District
Shanghai, 200433, China

First version: May 2009
This version: October 2016

Abstract

In this paper, we present a systematic study of partial identification of two general classes of functionals of the joint distribution of two "potential outcomes" when a bivariate sample from the joint distribution is not available to the econometrician. Assuming the identification of the conditional marginal distributions of potential outcomes and the distribution of the covariate vector, we show that the identified sets for functionals in both classes are intervals and provide conditions under which the identified sets point identify the true value of the functionals. In addition, we establish sufficient and necessary conditions for the covariate information to be informative in the sense of shrinking the identified sets. We focus on the application of our general results to evaluating distributional treatment effects of a binary treatment in two commonly used frameworks in the literature for evaluating average treatment effects: the selection on observables framework and a latent threshold-crossing model. We characterize the role of the propensity score in the selection-on-observables framework and the role of endogenous selection in the latent threshold-crossing model. Examples of policy parameters that our results apply include the correlation coefficient between the potential outcomes, many inequality measures of the distribution of treatment effects, and median of the distribution of the individual treatment effect.

Keywords: Copula; Distributional treatment effect; Selection-on-observables; Latent threshold-crossing model

JEL codes: C31, C14, C19, C39

*This is a substantially revised version of the previously circulated paper, Partial Identification and Confidence Sets for Functionals of the Joint Distribution of "Potential Outcomes". We thank Stephane Bonhomme, Yingyao Hu, Shih-Tang Hwu, Simon Lee, Konrad Menzel, Stephen Shore, Kevin Song, Richard Spady, Joerg Stoye, Tiemen Woutersen, participants of Bates White Sixth Annual Antitrust Conference 2009, Southern Economics Association Meetings 2009, International Symposium on Econometrics of Specification Tests in 30 Years at Xiamen University, 2010, and seminar participants at City University of Hong Kong, Johns Hopkins University, New York University, University of Kansas, Yale University, IUPUI, Emory University, Caltech, and Shanghai University of Finance and Economics for helpful comments and discussions on the previous paper.

1 Introduction

Parameters that depend on the joint distribution of two random variables are identified when a bivariate random sample from the joint distribution of the two variables is available. In many important applications in economics, finance, and other disciplines, however, such a bivariate random sample is not available. This paper considers this latter situation. Specifically let $Y_1 \in \mathcal{Y}_1$ and $Y_0 \in \mathcal{Y}_0$ denote two real-valued continuous random variables with joint cdf $F_o(y_1, y_0)$, $y_1 \in \mathcal{Y}_1$ and $y_0 \in \mathcal{Y}_0$. Let θ_o denote the parameter of interest. It is defined as[1] $\theta_o \equiv E_o[\mu(Y_1, Y_0)] \in \Theta \subset \mathcal{R}$ for some real-valued measurable function $\mu(\cdot,\cdot)$, where E_o denotes the expectation taken with respect to $F_o(\cdot,\cdot)$.

Assuming that the conditional marginal distributions of Y_1, Y_0 given a vector of covariates (which may contain unobserved components) and the distribution of the covariates are identified (Assumption (IC) in Section 3), this paper provides a systematic study of (partial) identification of θ_o for two general classes of functions μ. The first class is characterized by super-modular functions μ (see Definition 3.1) and the second by what we call φ-indicator functions ($\mu(Y_1, Y_0) \equiv I\{\varphi(Y_1, Y_0) \leq \delta\}$, see Definition 3.3 or Embrechts, Hoeing, and Puccetti (2005)). Building on existing works in the probability literature on solutions to the general Fréchet problem including a continuous version of the classical monotone rearrangement inequality,[2] this paper makes two original contributions. First, we characterize the identified sets for θ_o taking into account the covariate information for both classes of parameters and show that the identified set of the true parameter in each class is a closed interval. Second, for parameters corresponding to strict super-modular functions and parameters corresponding to φ functions that are strictly increasing in each argument, we establish sufficient and necessary conditions for point identification of the true parameter as well as sufficient and necessary conditions for the covariate to be informative in the sense of shrinking the identified set.

These general results have immediate applications in diverse areas including evaluation of distributional treatment effects where Y_1, Y_0 denote the potential outcomes of a binary treatment; bivariate option pricing where Y_1, Y_0 are prices of the underlying assets; and evaluation of the stop-loss premium of a portfolio of contracts. In this paper, we focus on their applications in the evaluation of distributional treatment effects and refer interested readers to the on-line Supplementary Appendices for examples in finance and insurance as well as related references.

Throughout the paper, we adopt two general frameworks in the treatment effect literature: the selection-on-observables framework and the latent threshold-crossing model in Heckman and Vytlacil (2005) and Carneiro and Lee (2009). Under commonly used assumptions in existing work to identify average treatment effects (ATE), both frameworks satisfy our Assumption (IC) and the general results established in this paper are applicable to both models. Examples of θ_o in the first class of parameters include the correlation coefficient between the potential outcomes, values of the joint distribution of the potential outcomes, and

[1] We'll introduce and discuss a conditional version of θ_o later in the paper.

[2] See Hardy, Littlewood, and Polya (1934), Cambanis, Simons, and Stout (1976), Tchen (1980), and Rachev and Ruschendorf (1998) for the first class of parameters; Makarov (1981), Rüschendorf (1982), and Frank, Nelsen, and Schweizer (1987), and Williamson and Downs (1990) for the second class of parameters.

many inequality measures of the distribution of treatment effects, see Examples (i) and (ii) in Section 2. Members of the second class of parameters include values of the cdf of treatment effects and quantiles of the distribution of treatment effects.[3] Heckman, Smith, and Clements (1997) and Abbring and Heckman (2007), among others, provide many examples demonstrating the need for evaluating joint distributions of potential outcomes, distributions of treatment effects, or other features of the distributions of treatment effects than various average treatment effects. Because of the missing data problem, evaluating these parameters is known to pose more challenges than evaluating average treatment effects, the latter being the focus of most works in the treatment effect literature, see Lee (2005), Abbring and Heckman (2007), Heckman and Vytlacil (2007a, b) for discussions and references. The current paper makes several contributions to the treatment effect literature.

First, it establishes identified sets for the afore-mentioned treatment effect parameters as well as sufficient and necessary conditions for their point identification in the context of selection-on-observables framework and latent threshold-crossing models. Second, in the selection-on-observables framework, we characterize the role of the propensity score and show that in sharp contrast to the identification of average treatment effects which can be based on either the observable covariates or the propensity score, the identified sets of distributional treatment effect parameters such as the correlation coefficient and the median of the distribution of treatment effects using the observable covariates could be tighter than the identified sets based on the propensity score. We provide sufficient and necessary conditions under which the two identified sets are the same. Third, we characterize the identified sets for distributional treatment effect parameters and the role of endogenous selection in the latent threshold-crossing model adopted in Heckman and Vytlacil (2005) and Carneiro and Lee (2009) to identify average treatment effect parameters. Fourth, to illustrate the important role played by the covariate (observable and unobservable), we provide a detailed analysis of the identified set of the correlation coefficient. In particular, we establish sufficient and necessary conditions for its identified set to exclude 0 when there is one observable covariate and when there is endogenous selection in the context of a latent threshold-crossing model. These conditions demonstrate clearly the role of the covariate information and endogenous selection in tightening the identified set. For ideal randomized experiments, Heckman, Smith, and Clements (1997) concluded that the bounds on the correlation coefficient between the potential outcomes implied by the result in Cambanis, Simons, and Stout (1976), i.e., without covariate, are often too wide to be informative. Our results show that (i) by exploiting information in the observable covariate, these bounds can be narrowed greatly and may be informative about the sign of the correlation coefficient when the dependence between the potential outcomes and the observable covariate is strong enough; and (ii) in the context of latent threshold-crossing model with endogenous selection, the requirement on the dependence between the potential outcomes and the observable covariate in (i) can be weakened significantly.

[3] Although quantiles of the distribution of treatment effects can not be written in the form of $\theta_o \equiv E_o\left[\mu\left(Y_1, Y_0\right)\right]$, their bounds follow immediately from bounds on the cdf of treatment effects and the cdf of the portfolios. So we simply refer to them as members of the second class of parameters.

This paper is related to several existing works on partial identification of treatment effects beyond the average treatment effect such as Manski (1997), Heckman, Smith, and Clements (1997), Fan and Park (2009, 2010, 2012), Fan and Wu (2010), Firpo and Ridder (2008), and Fan, Sherman, and Shum (2014). Assuming monotone treatment response, Manski (1997) developed sharp bounds[4] on the distributions of treatment effects; while assuming the availability of ideal randomized data, Heckman, Smith, and Clements (1997) used the result in Cambanis, Simons, and Stout (1976) to bound the correlation coefficient between the potential outcomes and the variance of the treatment effects. Fan, Sherman, and Shum (2014) examined partial identification of treatment effects under data combination.

Fan and Park (2009, 2010, 2012), Fan and Wu (2010), and Firpo and Ridder (2008) are the most closely related papers to the current paper. Besides studying a narrower class of parameters in Fan and Park (2009, 2010, 2012), they focus on ideal randomized experiments for which only the marginal cdfs of (Y_1, Y_0) are known (Assumption (I) in Section 3) or identified from the sample information. Within this framework, (i) Fan and Park (2009, 2010) study sharp bounds (pointwise) on the cdf of $\Delta = Y_1 - Y_0$ and their inference, from which they derive sharp bounds on the class of D-parameters including the quantile of the distribution of Δ and the class of D_2-parameters including Examples (i) and (ii) in the current paper; (ii) Fan and Park (2012) develop estimation and inference procedures for the quantile of the distribution of Δ. While Fan and Park (2009, 2010) briefly mentioned sharp bounds on the distribution of the treatment effect and their estimation under the selection-on-observables framework, they neither characterized its identified set nor investigated the role of the covariate in shrinking the identified set. In the context of switching regime models in Heckman (1990), Fan and Wu (2010) studied partial identification and (parametric) inference for conditional distributions of treatment effects given observable covariates.

Firpo and Ridder (2008) considered bounding a general functional of the distribution of treatment effects Δ. Note that the bounds on a general functional of the distribution of treatment effects obtained from the bounds on the distribution of treatment effects in Fan and Park (2009, 2010), and Firpo and Ridder (2008) are in general not sharp, as the bounds on the distribution of treatment effects are pointwise sharp, but not uniformly sharp. Firpo and Ridder (2008) presented a general approach to establishing bounds on functionals of the distribution of treatment effects that are tighter than bounds obtained directly from bounds on the distribution of treatment effects. However, the bounds in Firpo and Ridder (2008) are not sharp.

The rest of this paper is organized as follows. In Section 2, we first review the selection-on-observables framework and the latent threshold-crossing model in Heckman and Vytlacil (2005) and Carneiro and Lee (2009). Then we present some examples of the parameter θ_o measuring treatment effects beyond the ATE. In Section 3, we characterize the identified sets for the class of super-modular functions and of φ-indicator functions under Assumption (IC) and establish sufficient and necessary conditions for (i) the identified sets

[4] When we say bounds on a parameter θ_o, we mean a lower bound and an upper bound such that θ_o lies between the lower and upper bounds. When these bounds are achievable by some data generating process consistent with model assumptions, they are sharp bounds. These are terminologies used in the statistics and probability literature. In econometrics, we are interested in the identified set of a parameter. For example, the closed interval defined by the sharp bounds on a parameter is its identified set if each and every possible value in the interval can be realized for some data generating process consistent with model assumptions.

to be singleton and (ii) the covariate to shrink the identified sets. Section 4 examines the role of propensity score in the context of selection-on-observables framework and the role of endogenous selection in latent threshold-crossing models in shrinking the identified sets. Section 5 concludes and presents some extensions. Technical proofs are collected in Appendix A. Appendix B outlines an inference procedure for θ_o when μ is super-modular and the selection-on-observables assumption holds. Appendix C presents detailed derivations of the results discussed in Example (i)-(IC) in Section 3 and Example (i)-(IU) in Section 4. The on-line Supplementary Appendices contain additional examples, references, and technical proofs for the results in Appendix B in the current paper.

2 Identification of Treatment Effects With Observational Data

Let Y_1, Y_0 denote the potential outcomes of a binary treatment with an absolutely continuous joint cdf $F_o(y_1, y_0)$, $y_1 \in \mathcal{Y}_1$, $y_0 \in \mathcal{Y}_0$. Let $Y \equiv Y_1 D + Y_0(1-D)$ denote the realized outcome, where D is the binary treatment indicator such that an individual with $D = 1$ receives the treatment and an individual with $D = 0$ does not receive the treatment.

For an observable covariate X with support $\mathcal{X} \subset \mathcal{R}^d$, most treatment effect parameters of interest can be expressed as $\theta_o \equiv E_o[\mu(Y_1, Y_0)] \in \Theta \subset \mathcal{R}$ for some real-valued measurable function $\mu(\cdot,\cdot)$ or $\theta_o(x) \equiv E_o[\mu(Y_1, Y_0)|X = x]$ for $x \in \mathcal{X}$, where $E_o(\cdot)$ denotes the expectation taken with respect to $F_o(\cdot,\cdot)$ and $E_o(\cdot|X = x)$ denotes the expectation taken with respect to the conditional distribution of (Y_1, Y_0) given $X = x$. For example, the ATE and the conditional ATE correspond to $\mu(Y_1, Y_0) = Y_1 - Y_0$. As discussed in Heckman, Smith, and Clements (1997), many important policy questions can not be addressed by ATE parameters alone. Some examples and the corresponding functions μ are given in Subsection 2.2 below. In Subsection 2.1, we provide a brief review of the selection-on-observables framework and the latent threshold-crossing model in Heckman and Vytlacil (2005) and conditions under which ATEs are point identified in each framework.

2.1 The Selection-on-Observables Framework and a Latent Threshold-Crossing Model

The Selection-on-Observables Framework To identify various average treatment effect parameters, the selection-on-observables framework is commonly adopted in the literature, see e.g., Rosenbaum and Rubin (1983a, b), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998a, b), Dehejia and Wahba (1999), and Hirano, Imbens, and Ridder (2003), to name only a few. It is characterized by Assumption (IX) below.

Assumption (IX).

(C1) For all $x \in \mathcal{X} \subset \mathcal{R}^d$, (Y_1, Y_0) is jointly independent of D conditional on $X = x$.

(C2) For all $x \in \mathcal{X}$, $0 < p(x) < 1$, where $p(x) = \Pr(D = 1|X = x)$.

In Assumption (IX), (C1) is a conditional independence assumption and (C2) is a common support assumption. Suppose a random sample on (Y, X, D) is available. Then under Assumption (IX), for all $x \in \mathcal{X}$, the conditional marginal cdfs of Y_1, Y_0 given $X = x$ denoted as $F_{1o}(y|x)$ and $F_{0o}(y|x)$ are point identified:

$$F_{1o}(y|x) \equiv \Pr(Y_1 \leq y|X = x) = \Pr(Y \leq y|X = x, D = 1) \text{ and} \tag{1}$$

$$F_{0o}(y|x) \equiv \Pr(Y_0 \leq y|X = x) = \Pr(Y \leq y|X = x, D = 0). \tag{2}$$

Moreover, since the distribution of X is identified, the unconditional marginal cdfs $F_{1o}(y)$, $F_{0o}(y)$ are also point identified. As a result both ATE $E_o(\Delta)$ and the conditional ATE $E_o(\Delta|X = x)$ are point identified.

A Latent Threshold-Crossing Model Consider the semiparametric latent threshold-crossing model with continuous outcomes in Heckman (1990), Heckman and Vytlacil (1999, 2001, 2005):

$$Y_1 = g_1(X, U_1),\ Y_0 = g_0(X, U_0),\ \text{and } D = I\{g(Z) - \epsilon > 0\}, \tag{3}$$

where $X \in \mathcal{X} \subset \mathcal{R}^{d_x}$, $Z \in \mathcal{Z} \subset \mathcal{R}^{d_z}$ are observable covariates, U_1, U_0, ϵ are unobservable univariate covariates, g_1, g_0 and g are unknown functions, and the distribution of the unobserved error vector $(U_1, U_0, \epsilon)'$ is also unknown. Unlike the selection-on-observables framework, the latent threshold-crossing model in (3) allows for endogenous selection.

Suppose a random sample on (Y, X, Z, D) is available. Heckman and Vytlacil (2005) provided conditions under which various average treatment effect parameters are point identified, while Carneiro and Lee (2009) extended the results in Heckman and Vytlacil (2005) to the identification of distributions of $(Y_1, \epsilon)'$ and $(Y_0, \epsilon)'$ conditional on the observables. We restate these conditions in Assumption (IU) and Assumption (LS) below.

Assumption (IU). Assume that (i) $g(Z)$ is a nondegenerate random variable conditional on X; (ii) $(U_1, \epsilon)'$ and $(U_0, \epsilon)'$ are independent of Z conditional on X; (iii) the distribution of ϵ conditional on X, Z and that of $g(Z)$ conditional on X are absolutely continuous with respect to Lebesgue measure.

Without loss of generality, we normalize the distribution of ϵ conditional on X and Z to be $U(0, 1)$, implying by Assumption (IU)-(ii) that the distribution of ϵ conditional on X is also $U(0, 1)$. Let $p(z) = \Pr(D = 1|Z = z)$**.** Then $p(z) = g(z)$. Let $\mathcal{P}_x$ denote the support of $p(Z)$ conditional on $X = x \in \mathcal{X}$.

Assumption (LS). For each $x \in \mathcal{X}$, the closure of $\mathcal{P}_x$ is $[0, 1]$.

Let $X^* = (X', \epsilon)'$. It follows from Theorem 1 in Carneiro and Lee (2009) that under Assumptions (IU) and (LS), $F_{1o}(y|x^*)$ and $F_{0o}(y|x^*)$ are point identified from the sample information. In particular, they showed that

$$F_{1o}(y|x^*) = \Pr(Y \leq y|p(Z) = p, X = x, D = 1) + p\frac{\partial \Pr(Y \leq y|p(Z) = p, X = x, D = 1)}{\partial p} \text{ and} \tag{4}$$

$$F_{0o}(y|x^*) = \Pr(Y \leq y|p(Z) = p, X = x, D = 0) - (1 - p)\frac{\partial \Pr(Y \leq y|p(Z) = p, X = x, D = 0)}{\partial p}, \tag{5}$$

where $x^* = (x, p)$. Additionally, owing to the fact that the distribution of ϵ conditional on X is $U(0,1)$ (implying that the distribution of X^* is identified), it is easy to see that the unconditional marginal cdfs $F_{1o}(y)$, $F_{0o}(y)$ are also point identified from the sample information. Again both ATE and the conditional ATE are point identified.

Remark 2.1. Heckman and Vytlacil (1999, 2001, 2005) and Carneiro and Lee (2009) discuss in detail Assumptions (IU) and (LS). The main condition in Assumption (IU) is the exclusion restriction required to handle endogenous selection. Assumption (LS) is a large support restriction. When it fails, the conditional marginal cdfs may not be identified but may be bounded as in Heckman and Vytlacil (1999).

2.2 Treatment Effects Beyond ATE

This section provides examples of θ_o which measure other treatment effects than the ATE. In contrast to the ATE, these parameters depend not only on the marginal cdfs of Y_1, Y_0 but also their copula function. In all these examples, one can consider the conditional parameter $\theta_o(x)$ as well.

Example (i) (The Correlation Coefficient). Let $\mu(Y_1, Y_0) = Y_1 Y_0$ and $\sigma_j^2 = Var(Y_j) < \infty$ for $j = 0, 1$. Then the correlation coefficient between Y_1 and Y_0 is given by

$$\rho_{10} = \frac{E_o[\mu(Y_1, Y_0)] - E(Y_1) E(Y_0)}{\sigma_1 \sigma_0}.$$

Since $E(Y_j)$ and $Var(Y_j)$ depend on the marginal distributions only, we sometimes refer to $E_o[\mu(Y_1, Y_0)]$ as the correlation coefficient in which case $\mu(Y_1, Y_0) = Y_1 Y_0$.

Example (ii) (Distributional Treatment Effects I). Let $\Delta \equiv Y_1 - Y_0$ denote the individual treatment effect and $\mu(Y_1, Y_0) = \nu(\Delta)$ for some function ν. Many inequality measures of the distribution of treatment effect Δ can be expressed as $g(E_o[\nu(\Delta)], \mu_\Delta)$, where $\mu_\Delta \equiv E_o(\Delta)$ is the ATE, $g(\cdot, \cdot)$ is increasing in its first argument, and $\nu(\cdot)$ is continuous and convex, see Stoye (2010) and references therein. For instance, the coefficient of variation defined as

$$\theta_{CV} = \frac{\sqrt{Var_o(\Delta)}}{\mu_\Delta} = \frac{\sqrt{E_o(\Delta^2) - \mu_\Delta^2}}{\mu_\Delta}$$

can be written as $g(E_o[\nu(\Delta)], \mu_\Delta)$, where $\nu(\Delta) = \Delta^2$ is continuous and convex and $g(z, \mu_\Delta) = \sqrt{z - \mu_\Delta^2}/\mu_\Delta$ is increasing in z. A general class of inequality measures of the distribution of Δ is that of generalized entropy measures. Let γ denote an even number, $\nu_\gamma(\Delta) = \Delta^\gamma$, and

$$g_\gamma(z, \mu_\Delta) = \frac{1}{\gamma^2 - \gamma}\left[\frac{z}{\mu_\Delta^\gamma} - 1\right].$$

Then $\nu_\gamma(\cdot)$ is continuous and convex. Further $g_\gamma(E_o[\nu_\gamma(\Delta)], \mu_\Delta)$ is a generalized entropy measure of the distribution of Δ.

Example (iii) (Distributional Treatment Effects II). (a) Let $\mu(Y_1, Y_0) = 1(\Delta > 0)$. The proportion of people who benefit from the treatment is given by

$$E_o[\mu(Y_1, Y_0)] = \Pr(\Delta > 0) = 1 - F_\Delta(0),$$

where $F_\Delta(\cdot)$ is the cdf of Δ. (b) Let $\alpha \in (0,1)$. Although the α-quantile of the distribution of Δ, $F_\Delta^{-1}(\alpha)$, is strictly speaking not an example of θ_o, its bounds can be obtained by inverting the bounds on $F_\Delta(\delta) = E_o[\mu(Y_1, Y_0)]$ with $\mu(Y_1, Y_0) = 1(\Delta \leq \delta)$, and thus we simply refer to $F_\Delta^{-1}(\alpha)$ as an example of θ_o.

Throughout the rest of this paper, we adopt either the selection-on-observables framework or the latent threshold-crossing model satisfying Assumptions (IU) and (LS). In either case, the conditional marginal cdfs of (Y_1, Y_0) given X^* and the cdf of X^* are point identified, where X^* is observable in the former case and contains an unobservable component in the latter model. Although ATE and the conditional ATE are point identified in both frameworks, parameters in Examples (i)-(iii) and their conditional versions are not point identified without further assumptions. This paper characterizes their identified sets.

3 Partial Identification of Treatment Effects Beyond ATE

This section provides a unified analysis of identification of $\theta_o \equiv E_o[\mu(Y_1, Y_0)]$ under Assumption (IC) below in which $X^* \in \mathcal{X}^* \subset \mathcal{R}^{d^*}$ denotes the vector of covariates which may contain unobservable components. The corresponding analysis for the conditional parameter $\theta_o(x) = E_o[\mu(Y_1, Y_0)|X = x]$, where X is the observed component of X^*, is discussed in Remarks 3.1 and 3.2 for super-modular and φ-indicator functions respectively.

Assumption (IC). The conditional marginal cdfs of Y_1, Y_0 given $X^* = x^*$ denoted as $F_{1o}(\cdot|x^*)$ and $F_{0o}(\cdot|x^*)$ are known for all $x^* \in \mathcal{X}^*$. Moreover the cdf of X^* denoted as $F_{X^*o}(\cdot)$ is also known.

In stating Assumption (IC) and Assumption (I) below, we have followed the tradition in the literature on identification by referring to $F_{1o}(\cdot|x^*)$, $F_{0o}(\cdot|x^*)$, $F_{X^*o}(\cdot)$, $F_{1o}(\cdot)$, and $F_{0o}(\cdot)$ as known. In specific applications, they are point identified from model assumptions and the sample information such as in the selection-on-observables framework and latent threshold-crossing models reviewed in Section 2.

Let $C_o(\cdot, \cdot|x^*)$ denote the conditional copula of Y_1, Y_0 given $X^* = x^*$, where $x^* \in \mathcal{X}^*$. We note that θ_o can be expressed as the following form:

$$\begin{aligned}\theta_o &= E\left\{\int\int \mu(y_1, y_0)\, dF_o(y_1, y_0|X^*)\right\} \\ &= E\left\{\int\int \mu(y_1, y_0)\, dC_o(F_{1o}(y_1|X^*), F_{0o}(y_0|X^*)|X^*)\right\}.\end{aligned}$$

Under Assumption (IC), the identified set for θ_o is given by

$$\Theta_{IC} \equiv \left\{ \begin{array}{c} \theta \in \Theta : \theta = E\left[\int\int \mu(y_1, y_0)\, dC(F_{1o}(y_1|X^*), F_{0o}(y_0|X^*)|X^*)\right] \\ \text{for some } C(\cdot, \cdot|X^*) \in \mathcal{C} \text{ a.s.} \end{array} \right\}, \tag{6}$$

where $\mathcal{C}$ denotes the class of bivariate copula functions.

Existing works such as Heckman, Smith, and Clements (1997) and Fan and Park (2009, 2010, 2012) studied specific examples of θ_o under Assumption (I) below.

Assumption (I). The marginal cdfs of Y_1, Y_0 denoted as $F_{1o}(\cdot)$ and $F_{0o}(\cdot)$ are known.

Under Assumption (I), the identified set for θ_o is given by:

$$\Theta_I \equiv \left\{ \begin{array}{c} \theta \in \Theta : \theta = E\left[\int\int \mu(y_1, y_0)\, dC(F_{1o}(y_1), F_{0o}(y_0))\right] \\ \text{for some } C(\cdot, \cdot) \in \mathcal{C} \end{array} \right\}. \tag{7}$$

The difference between the two identified sets Θ_{IC} and Θ_I reflects the role played by the covariate X^* in shrinking the identified set of θ_o, since Assumption (IC) implies that the marginal cdfs $F_{1o}(\cdot)$ and $F_{0o}(\cdot)$ are known. In the rest of this section, we first characterize the identified set Θ_{IC} for super-modular functions and φ-indicator functions. Then for strict super-modular and φ-indicator functions with φ being strictly monotone in each argument, we establish necessary and sufficient conditions for (i) Θ_{IC} or Θ_I to be a singleton thus point identifying θ_o and (ii) for X^* to be informative in the sense of shrinking the identified set for θ_o, i.e., for Θ_{IC} to be smaller than Θ_I.

3.1 A Characterization of Θ_{IC} for Super-Modular Functions and the Role of the Covariate

We first present a definition of a super-modular[5] function.

Definition 3.1 *A function $\mu(\cdot,\cdot)$ is called super-modular if for all $y_1 \leq y_1'$ and $y_0 \leq y_0'$,*

$$\mu(y_1, y_0) + \mu(y_1', y_0') - \mu(y_1, y_0') - \mu(y_1', y_0) \geq 0,$$

and sub-modular if $-\mu(\cdot,\cdot)$ is super-modular.

If $\mu(\cdot,\cdot)$ is absolutely continuous, then it is super-modular if and only if $\frac{\partial^2 \mu(y_1,y_0)}{\partial y_1 \partial y_0} \geq 0$ a.e. Cambanis, Simons, and Stout (1976) provide many examples of super-modular or sub-modular functions, see also Tchen (1980). The μ functions in Examples (i) and (ii) are either super-modular or sub-modular. The difference function $\mu(y_1, y_0) = y_1 - y_0$ is also super-modular. Other examples of super-modular or sub-modular functions include: $\mu(y_1, y_0) = (\min(y_1, y_0) - k)_+$, $\mu(y_1, y_0) = (y_1 + y_0 - k)_+$ for some known number k, where $(x)_+ = \max(x, 0)$ and $\mu(y_1, y_0) = \min\left\{(y_1 - k_1)_+, (y_0 - k_0)_+\right\}$ for some known values k_1, k_0. These are payoff functions of specific bivariate options, see the on-line Supplementary Appendices for details.

The function $\mu(y_1, y_0) = y_1 - y_0$ is different from the other functions above in that it is additively separable in its arguments. The μ functions in Examples (i) and (ii) belong to the class of strict super-modular or strict sub-modular functions defined below.

Definition 3.2 *A function $\mu(\cdot,\cdot)$ is called "strict super-modular" if it is super-modular and for all $y_1 < y_1'$ and $y_0 < y_0'$, it holds that*

$$\mu(y_1, y_0) + \mu(y_1', y_0') - \mu(y_1, y_0') - \mu(y_1', y_0) > 0,$$

and strict sub-modular if $-\mu(\cdot,\cdot)$ is strict super-modular.

It is clear that a strict super-modular or strict sub-modular function can not be additively separable in its arguments, but a super-modular or sub-modular function can. Other examples of additively separable super-modular functions include $\mu(y_1, y_0) = h_1(y_1) - h_0(y_0)$ for known measurable functions h_1 and h_0, see

[5] A super-modular function is also called a quasi-monotone function or a super-additive function in probability and statistics literature.

Firpo and Pinto (2015) for measures of treatment effects corresponding to such functions μ. In addition to the μ functions in Examples (i) and (ii), other examples of strict super-modular or sub-modular functions $\mu(\cdot,\cdot)$ include $\mu(y_1,y_0) = h_1(y_1)h_0(y_0)$, where h_1 and h_0 are known strictly monotonic functions. For example, Spearman's rank correlation, $\rho_S(Y_1,Y_0) \equiv corr[F_{1o}(Y_1), F_{0o}(Y_0)]$, corresponds to $\mu(y_1,y_0) = F_{1o}(y_1)F_{0o}(y_0)$.

3.1.1 Some Basic Results on Θ_I

For a super-modular and right continuous function $\mu(\cdot,\cdot)$ satisfying some regularity conditions, the identified set for θ_o is a closed interval, see e.g., Cambanis, Simons, and Stout (1976), Tchen (1980), and Rachev and Ruschendorf (1998).[6] To introduce it, let

$$F^{(-)}(y_1,y_0) \equiv M(F_{1o}(y_1), F_{0o}(y_0)) \text{ and } F^{(+)}(y_1,y_0) \equiv W(F_{1o}(y_1), F_{0o}(y_0)),$$

where $M(u,v) \equiv \max(u+v-1,0)$ and $W(u,v) \equiv \min(u,v)$ are the Fréchet-Hoeffding lower and upper bounds for a copula. Then $\Theta_I = \left[\theta^L, \theta^U\right]$, where

$$\begin{aligned}\theta^L &\equiv E_{F^{(-)}}[\mu(Y_1,Y_0)] = \int_0^1 \mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(1-u)\right) du \text{ and} \\ \theta^U &\equiv E_{F^{(+)}}[\mu(Y_1,Y_0)] = \int_0^1 \mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) du, \end{aligned} \tag{8}$$

in which E_F denotes the expectation taken with respect to the joint cdf F and $F_{jo}^{-1}(u) = \inf\{y : F_{jo}(y) \geq u\}$ is the quantile function of Y_j, $j = 0, 1$.

If $\mu(\cdot,\cdot)$ is additively separable in its arguments, then $\theta^L = \theta^U$ in which case θ_o is point identified for all the marginal distribution functions F_{1o}, F_{0o} under Assumption (I). However, when $\mu(\cdot,\cdot)$ is not additively separable in its arguments, in general $\theta^L \neq \theta^U$ and θ_o is only partially identified. Below we show that under the conditions of Theorem 2 in Cambanis, Simons, and Stout (1976) restated as conditions (a) and (b) in Theorem 3.1, for strict super-modular functions $\mu(\cdot,\cdot)$, θ_o is point identified only in trivial cases, i.e., when at least one of the marginal distributions F_{1o}, F_{0o} is degenerate.

THEOREM 3.1 *Suppose that Assumption (I) holds and let $\mu(y_1,y_0)$ be a super-modular and right continuous function. Suppose that θ^L and θ^U exist (even if infinite valued) and that either of the following conditions is satisfied: (a) $\mu(y_1,y_0)$ is symmetric and $E[\mu(Y_1,Y_1)]$ and $E[\mu(Y_0,Y_0)]$ are finite (in this case, $-\infty \leq \theta^L \leq \theta^U < +\infty$); (b) there are some fixed constants $\overline{y}_0$ and $\overline{y}_1$ such that $E[\mu(Y_1,\overline{y}_0)]$ and $E[\mu(\overline{y}_1,Y_0)]$ are finite and at least one of θ^L and θ^U is finite. Then (i) when $\mu(\cdot,\cdot)$ is additively separable, $\theta^L = \theta^U$; (ii) when $\mu(\cdot,\cdot)$ is strict super-modular, $\theta^L = \theta^U$ if and only if at least one of the marginal distributions F_{1o}, F_{0o} is degenerate.*

Following discussions in Cambanis, Simons, and Stout (1976), it is clear that when random variables Y_1, Y_0 are bounded, condition (b) is satisfied for locally bounded functions $\mu(y_1,y_0)$ such as those in Examples (i)

[6] Results for sub-modular functions follow straightforwardly from the corresponding results for super-modular functions. To save space, we will not present results for sub-modular functions in this paper.

and (ii) and payoff functions of specific bivariate options mentioned above, and if $\mu(y_1, y_0)$ is also symmetric, then condition (a) is also satisfied.

3.1.2 A Characterization of Θ_{IC} and the Role of the Covariate

Let

$$\theta_L \equiv E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X^*), F_{0o}^{-1}(1-u|X^*)\right) du\right] \text{ and} \tag{9}$$
$$\theta_U \equiv E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X^*), F_{0o}^{-1}(u|X^*)\right) du\right],$$

where $F_{jo}^{-1}(u|x^*) = \inf\{y : F_{jo}(y|x^*) \geq u\}$ is the quantile function of Y_j conditional on $X^* = x^*$, $j = 0, 1$. Theorem 3.2 below extends Theorem 2 in Cambanis, Simons, and Stout (1976) and Theorem 3.1 above, characterizing the identified set for θ_o under Assumption (IC) for super-modular and right continuous functions μ.

THEOREM 3.2 *Suppose that Assumption (IC) holds and let $\mu(y_1, y_0)$ be a super-modular and right continuous function. Suppose that both expectations in (9) exist (even if infinite valued) and that either of the conditions (a) and (b) in Theorem 3.1 (with θ_L and θ_U replacing θ^L and θ^U in condition (b)) is satisfied. Then*

(i) the identified set for $\theta_o = E_o[\mu(Y_1, Y_0)]$ is $\Theta_{IC} = [\theta_L, \theta_U]$;

(ii) for a strict super-modular function $\mu(\cdot,\cdot)$, $\theta_L = \theta_U$ if and only if at least one of the conditional marginal distributions $F_{1o}(\cdot|x^*), F_{0o}(\cdot|x^*)$ is degenerate for almost all $x^* \in \mathcal{X}^*$.

Note that under Assumption (IC), the joint cdfs of (Y_1, X^*) and (Y_0, X^*) are known and we expect the covariate X^* to contain information on the dependence between Y_1 and Y_0. As a result, compared with $\Theta_I = [\theta^L, \theta^U]$, the identified set $\Theta_{IC} = [\theta_L, \theta_U]$ should be shrunk. To show that $\Theta_{IC} \subseteq \Theta_I$, we let[7]

$$F_*^{(-)}(y_1, y_0) \equiv E[M(F_{1o}(y_1|X^*), F_{0o}(y_0|X^*))] \text{ and}$$
$$F_*^{(+)}(y_1, y_0) \equiv E[W(F_{1o}(y_1|X^*), F_{0o}(y_0|X^*))].$$

We can prove that θ_L and θ_U are attained when (Y_1, Y_0) has the cdfs $F_*^{(-)}(y_1, y_0)$ and $F_*^{(+)}(y_1, y_0)$ respectively and that the identified set $\Theta_{IC} = [\theta_L, \theta_U]$ is identical to the set of values of $E_F[\mu(Y_1, Y_0)]$ when F ranges over the class of all joint cdfs $F(\cdot,\cdot)$ with fixed marginals F_{1o} and F_{0o} satisfying $F_*^{(-)}(y_1, y_0) \leq F(y_1, y_0) \leq F_*^{(+)}(y_1, y_0)$. In other words,

$$\Theta_{IC} = \left\{ \begin{array}{c} \theta \in \Theta : \theta = E\left[\int\int \mu(y_1, y_0)\, dC(F_{1o}(y_1), F_{0o}(y_0))\right] \\ \text{for some } C(\cdot,\cdot) \text{ satisfying } M^*(\cdot,\cdot) \leq C(\cdot,\cdot) \leq W^*(\cdot,\cdot) \end{array} \right\}, \tag{10}$$

where $M^*(\cdot,\cdot)$ and $W^*(\cdot,\cdot)$ are defined as the copulas of the cdfs $F_*^{(-)}(y_1, y_0)$ and $F_*^{(+)}(y_1, y_0)$ respectively. Since $M(\cdot,\cdot) \leq M^*(\cdot,\cdot)$ and $W^*(\cdot,\cdot) \leq W(\cdot,\cdot)$, it holds that $\Theta_{IC} \subseteq \Theta_I$.

[7] Measurability of $M(F_{1o}(y_1|X^*), F_{0o}(y_0|X^*))$ and $W(F_{1o}(y_1|X^*), F_{0o}(y_0|X^*))$ follows from measurability of $F_{1o}(y_1|X^*)$ and $F_{0o}(y_0|X^*)$.

For strict super-modular functions μ, Theorem 3.3 below establishes sufficient and necessary conditions for $\Theta_{IC} = \Theta_I$ or equivalently for Θ_{IC} to be a proper subset of Θ_I.

THEOREM 3.3 *Suppose that Assumption (IC) holds and let $\mu(y_1, y_0)$ be a super-modular and right continuous function. Suppose that the four expectations in (8) and (9) exist (even if infinite valued) and that either of the conditions (a) and (b) in Theorem 3.1 is satisfied. Then $\Theta_{IC} \subseteq \Theta_I$ and if $\mu(\cdot,\cdot)$ is strict super-modular, then $\Theta_{IC} = \Theta_I$ iff for μ_c-almost all (y_1, y_0),[8] it holds that*

$$\Pr\left(F_{1o}(y_1|X^*) + F_{0o}(y_0|X^*) - 1 > 0\right) \in \{0, 1\} \text{ and} \tag{11}$$

$$\Pr\left(F_{1o}(y_1|X^*) - F_{0o}(y_0|X^*) < 0\right) \in \{0, 1\}. \tag{12}$$

Obviously, (11) and (12) hold if both Y_1 and Y_0 are independent of X^* in which case the covariate X^* does not help shrink the identified set Θ_I. Also if one of Y_1 and Y_0, say Y_1, is degenerate, then (11) and (12) hold, so $\Theta_I = \Theta_{IC}$, but both sets are singleton. When both Y_1 and Y_0 are not degenerate but for almost every $x^* \in \mathcal{X}^*$, at least one of the conditional marginal distributions $F_{1o}(\cdot|x^*)$ and $F_{0o}(\cdot|x^*)$ is degenerate, then (11) and (12) will not hold and in this case Θ_{IC} is singleton but Θ_I is not. For conditional distributions $F_{1o}(\cdot|x^*)$ and $F_{0o}(\cdot|x^*)$ that violate either (11) or (12), the identified set Θ_{IC} is a proper subset of Θ_I, so incorporating information in X^* helps shrink the identified set Θ_I. This can be useful when the identified set Θ_I is itself not informative as we show in Example (i)-(IC)[9] below for the correlation coefficient between Y_1 and Y_0.

Example (i)-(IC) (Correlation Coefficient). Let the covariate X^* be univariate. For notational simplicity, we denote X^* as X in this example. Suppose the distribution of (Y_j, X) is known to be a bivariate normal distribution:

$$\begin{pmatrix} Y_j \\ X \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_j^2 & \sigma_j\rho_{jX} \\ \sigma_j\rho_{jX} & 1 \end{pmatrix}\right], \; j = 0, 1.$$

Then Assumption (IC) is satisfied with $Y_j|X = x \sim N\left(\sigma_j\rho_{jX}x, \sigma_j^2\left(1-\rho_{jX}^2\right)\right)$, $j = 0, 1$, and $X \sim N(0, 1)$. Obviously, $Y_j \sim N\left(0, \sigma_j^2\right)$. Suppose $\sigma_j^2 > 0$, $j = 1, 0$. It is known that the identified set for ρ_{10} (i.e., the correlation coefficient between Y_1 and Y_0) under Assumption (I) is $\Theta_I = \left[\rho^L, \rho^U\right] = [-1, 1]$, see also Appendix C. It can not identify the sign of ρ_{10}. In Appendix C, we show that under Assumption (IC), the identified set $\Theta_{IC} = [\rho_L, \rho_U]$, where

$$\rho_L = \rho_{0X}\rho_{1X} - \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)} \text{ and } \rho_U = \rho_{0X}\rho_{1X} + \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)},$$

which identifies the sign of ρ_{10} as long as the dependence between (Y_1, Y_0) and covariate X is strong enough in the sense that $\rho_{0X}^2 + \rho_{1X}^2 > 1$ in which ρ_{jX} is the correlation coefficient between Y_j and X, $j = 1, 0$. In addition, Example (i)-(IC) validates Theorem 3.3 and Theorem 3.2 (ii).

[8] If $\mu(\cdot,\cdot)$ is super-modular and right continuous, then it uniquely determines a nonnegative measure μ_c on the Borel subsets of the plane $\mathcal{R}^2$ such that for all $y_1 \leq y_1'$ and $y_0 \leq y_0'$, $\mu_c\left(\left(y_1, y_1'\right] \times \left(y_0, y_0'\right]\right) = \mu(y_1, y_0) + \mu\left(y_1', y_0'\right) - \mu\left(y_1, y_0'\right) - \mu\left(y_1', y_0\right)$. See Cambanis, Simons, and Stout (1976), and Rachev and Ruschendorf (1998).

[9] The on-line Supplementary Appendices offer another example.

Remark 3.1. The results in Subsection 3.1.2 apply directly to θ_o under the two frameworks reviewed in Section 2, i.e., the selection-on-observables framework and the latent threshold-crossing model. When the parameter of interest is $\theta_o(x)$ for a given $x \in \mathcal{X}$, the identified sets take different forms under the selection-on-observables framework and the threshold-crossing model. Under the selection-on-observables assumption, the identified set for $\theta_o(x)$ is given by

$$\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|x), F_{0o}^{-1}(1-u|x)\right) du, \ \int_0^1 \mu\left(F_{1o}^{-1}(u|x), F_{0o}^{-1}(u|x)\right) du\right];$$

while in the threshold-crossing model, a straightforward extension of the argument for Theorem 3.2 (i) shows that it is the closed interval with end points given by

$$\int_0^1 \int_0^1 \mu\left(F_{1o}^{-1}(u|x,\epsilon), F_{0o}^{-1}(1-u|x,\epsilon)\right) du d\epsilon \text{ and}$$
$$\int_0^1 \int_0^1 \mu\left(F_{1o}^{-1}(u|x,\epsilon), F_{0o}^{-1}(u|x,\epsilon)\right) du d\epsilon,$$

where we used the fact that the distribution of ϵ conditional on X is $U(0,1)$.

3.2 A Characterization of Θ_{IC} for φ-Indicator Functions and the Role of the Covariate

Definition 3.3 *Let φ denote a measurable function and $\mu(Y_1, Y_0) \equiv I\{\varphi(Y_1, Y_0) \leq \delta\}$ for a fixed δ in the support of $\varphi(Y_1, Y_0)$. Moreover $\varphi(\cdot,\cdot)$ is monotone in each argument. We refer to this class of functions μ as the class of φ-indicator functions.*

Let $F_\varphi(\cdot)$ denote the distribution function of $\varphi(Y_1, Y_0)$. Then for a fixed δ, $\theta_o = \Pr(\varphi(Y_1, Y_0) \leq \delta) = F_\varphi(\delta)$. Building on the sharp bounds established in Frank, Nelsen, and Schweizer (1987), Williamson and Downs (1990), and Embrechts, Hoeing, and Juri (2003),[10] one can show that for the class of functions $\varphi(\cdot,\cdot)$ that are continuous and non-decreasing in each argument,[11] the identified set for θ_o under Assumption (I) is the closed interval with end points $F_{\min,\varphi}(\delta)$ and $F_{\max,\varphi}(\delta)$, i.e., $\Theta_I = [F_{\min,\varphi}(\delta), F_{\max,\varphi}(\delta)]$, where

$$F_{\min,\varphi}(\delta) = \sup_{y \in \mathcal{Y}_1} \max(F_{1o}(y) + F_{0o}(\hat{\varphi_y}(\delta)) - 1, 0) \text{ and}$$
$$F_{\max,\varphi}(\delta) = 1 + \inf_{y \in \mathcal{Y}_1} \min(F_{1o}(y) + F_{0o}(\hat{\varphi_y}(\delta)) - 1, 0),$$

in which $\hat{\varphi_y}(\delta) \equiv \sup\{y_0 \in \mathcal{Y}_0 : \varphi(y, y_0) < \delta\}$.

Making use of the above result for $\varphi(Y_1, Y_0) = Y_1 - Y_0$, Fan and Park (2009, 2010) provide a systematic study of partial identification and inference for $\theta_o = F_\Delta(\delta)$, while Fan and Park (2012) construct inference procedures for $F_\Delta^{-1}(\alpha)$ in ideal randomized experiments.

[10] See Frank, Nelsen, and Schweizer (1987) for sharp bounds for the sum of two random variables, Williamson and Downs (1990) for the four basic arithmetic operations, and Theorem 5.1 in Frank, Nelsen, and Schweizer (1987) and Embrechts, Hoeing, and Juri (2003) for general non-decreasing functions.

[11] Without loss of generality, we focus on the class of functions $\varphi(\cdot,\cdot)$ that are non-decreasing in each argument. The results obtained can be applied to other types of monotone functions $\varphi(\cdot,\cdot)$ by redefining either Y_1 or Y_0 appropriately. For example, when $\varphi(Y_1, Y_0) = Y_1 - Y_0$ which is decreasing in Y_0, we redefine the two random variables as Y_1 and $(-Y_0)$ to obtain a new function which is increasing in both arguments.

For a large class of functions φ, Theorem 3.4 below gives sufficient and necessary conditions for $F_{\varphi}(\delta)$ to be point identified.

THEOREM 3.4 *Suppose that φ is continuous and strictly increasing in each argument. Then $\Theta_I = [F_{\min,\varphi}(\delta), F_{\max,\varphi}(\delta)]$ is a singleton for all δ if and only if at least one of the unconditional marginal distributions F_{1o}, F_{0o} is degenerate.*

Now consider the identified set for θ_o under Assumption (IC). Let $\mathcal{Y}_1(X^*)$ and $\mathcal{Y}_0(X^*)$ be the supports of Y_1 and Y_0 given X^*, respectively, and define

$$\begin{aligned} F_{\min,\varphi}(\delta|X^*) &= \sup_{y\in\mathcal{Y}_1(X^*)} \max\{F_{1o}(y|X^*) + F_{0o}(\hat{\varphi_y}(\delta|X^*)|X^*) - 1, 0\} \text{ and} \\ F_{\max,\varphi}(\delta|X^*) &= 1 + \inf_{y\in\mathcal{Y}_1(X^*)} \min\{F_{1o}(y|X^*) + F_{0o}(\hat{\varphi_y}(\delta|X^*)|X^*) - 1, 0\}, \end{aligned} \tag{13}$$

where $\hat{\varphi_y}(\delta|X^*) = \sup\{y_0 \in \mathcal{Y}_0(X^*) : \varphi(y, y_0) < \delta\}$. Note that for a fixed δ in the support of $\varphi(Y_1, Y_0)$, the set $\{y_0 \in \mathcal{Y}_0(x^*) : \varphi(y, y_0) < \delta\}$ for some $y \in \mathcal{Y}_1(x^*)$ and $x^* \in \mathcal{X}^*$ may be empty. If so, $\hat{\varphi_y}(\delta|x^*)$ is defined as minus infinity. Theorem 3.5 below extends Theorem 1 in Williamson and Downs (1990) and Theorem 3.4 above, see also Embrechts, Hoeing, and Juri (2003).

THEOREM 3.5 *Suppose that Assumption (IC) holds and that φ is continuous and non-decreasing in each argument. Suppose that both $\mathcal{Y}_1(X^*)$ and $\mathcal{Y}_0(X^*)$ are the Borel sets generated by intervals with both ends being measurable. Then (i) the identified set for $\theta_o = F_{\varphi}(\delta)$ is $\Theta_{IC} = [F_{L,\varphi}(\delta), F_{U,\varphi}(\delta)]$, where $F_{L,\varphi}(\delta) = E[F_{\min,\varphi}(\delta|X^*)]$ and $F_{U,\varphi}(\delta) = E[F_{\max,\varphi}(\delta|X^*)]$; (ii) if φ is strictly increasing in each argument, then $F_{L,\varphi}(\delta) = F_{\varphi}(\delta) = F_{U,\varphi}(\delta)$ for all δ if and only if for almost every $x^* \in \mathcal{X}^*$, at least one of the conditional marginal distributions $F_{1o}(\cdot|x^*), F_{0o}(\cdot|x^*)$ is degenerate.*

Obviously supports $\mathcal{Y}_j(X^*)$ that are given by intervals with both ends being measurable satisfy the conditions of Theorem 3.5. Theorem 3.5 implies that for $0 < \alpha < 1$, $F^{-1}_{U,\varphi}(\alpha) \leq F^{-1}_{\varphi}(\alpha) \leq F^{-1}_{L,\varphi}(\alpha)$ extending the bounds on quantiles of treatment effects in Fan and Park (2012) for ideal randomized experiments.

Similar to Theorem 3.3 for super-modular functions, it is possible to establish conditions under which Θ_{IC} is a proper subset of Θ_I. To simplify the technical argument, Theorem 3.6 below provides such a result for the case[12] that $\mathcal{Y}_j(x^*) = \mathcal{Y}_j$ for $j = 0, 1$ and all $x^* \in \mathcal{X}^*$. In this case, $\hat{\varphi_y}(\delta|X^*) = \hat{\varphi_y}(\delta)$ with probability one.

THEOREM 3.6 *Suppose that the conditions of Theorem 3.5 hold and that $\mathcal{Y}_j(x^*) = \mathcal{Y}_j$ for $j = 0, 1$ and all $x^* \in \mathcal{X}^*$. Suppose that $\left[F_{1o}(y) + F_{0o}(\hat{\varphi_y}(\delta)) - 1\right]$ achieves the maximum and the minimum values at some $\overline{y} \in \mathcal{Y}_1$ and $\underline{y} \in \mathcal{Y}_1$, respectively. Then*

$$[F_{L,\varphi}(\delta), F_{U,\varphi}(\delta)] = [F_{\min,\varphi}(\delta), F_{\max,\varphi}(\delta)]$$

[12] By following the proof of Theorem 3.6, one can show that without the condition: $\mathcal{Y}_j(x^*) = \mathcal{Y}_j$ for $j = 0, 1$ and all $x^* \in \mathcal{X}^*$, the stated condition in Theorem 3.6 is still sufficient but whether it is still necessary needs to be investigated.

if and only if $\left[F_{1o}(y|x^*) + F_{0o}(\hat{\varphi_y}(\delta)|x^*) - 1\right]$ *achieves the maximum and the minimum values uniformly at* $\overline{y}$ *and* $\underline{y}$ *for almost all* $x^* \in \mathcal{X}^*$*, respectively.*

One sufficient condition for the 'iff' condition in Theorem 3.6 is that X^* is independent of (Y_1, Y_0). This implies that in general using covariates may shrink the identified set for $F_\varphi(\delta)$. Theorem 3.5 (ii) provides an example demonstrating the importance of this improvement. It says that when at least one of the potential outcomes is a deterministic function of X^*, the identified set $[F_{L,\varphi}(\delta), F_{U,\varphi}(\delta)]$ is a singleton and point identifies the parameter $F_\varphi(\delta)$. However, the interval $[F_{\min,\varphi}(\delta), F_{\max,\varphi}(\delta)]$ does not identify $F_\varphi(\delta)$ except in the trivial case where one of the potential outcomes is a constant with probability one, see Theorem 3.4.

Remark 3.2. Theorem 3.5 applies directly to θ_o under the two frameworks reviewed in Section 2, i.e., the selection-on-observables framework and the latent threshold-crossing model. When the parameter of interest is $\theta_o(x)$ for a given $x \in \mathcal{X}$, the identified sets take different forms under the selection-on-observables framework and the threshold-crossing model. Under the selection-on-observables assumption, the identified set for $\theta_o(x)$ is the closed interval $[F_{\min,\varphi}(\delta|X=x), F_{\max,\varphi}(\delta|X=x)]$, where $F_{\min,\varphi}(\delta|X)$ and $F_{\max,\varphi}(\delta|X)$ are defined in (13) with X^* replaced by X. In the threshold-crossing model, a straightforward extension of the argument for Theorem 3.5 (i) shows that it is the closed interval $\left[\int_0^1 F_{\min,\varphi}(\delta|x,\epsilon)d\epsilon, \int_0^1 F_{\max,\varphi}(\delta|x,\epsilon)d\epsilon\right]$.

4 The Role of the Propensity Score and the Role of Endogenous Selection

In the selection-on-observables framework, Rosenbaum and Rubin (1983a, b) show that (Y_1, Y_0) is also jointly independent of D conditional on the propensity score $p(X)$, so the average treatment effect can be point identified via conditioning on either X or $p(X)$:

$$\begin{aligned}\mu_\Delta &= E\left(E\left[Y_1|X, D=1\right] - E\left[Y_0|X, D=0\right]\right) \\ &= E\left(E\left[Y_1|p(X), D=1\right] - E\left[Y_0|p(X), D=0\right]\right).\end{aligned}$$

In contrast, for strict super-modular and right continuous functions or φ-indicator functions, the identified set based on the propensity score $p(X)$ may be larger than the identified set based on X. Proposition 4.1 below establishes a sufficient and necessary condition[13] under which the two sets are identical for strict super-modular functions.[14]

Suppose μ is super-modular and right continuous. Using the propensity score, we get the identified set $[\theta_{LP}, \theta_{UP}]$, where

$$\begin{aligned}\theta_{LP} &= E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|p(X)), F_{0o}^{-1}(1-u|p(X))\right) du\right] \text{ and} \\ \theta_{UP} &= E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|p(X)), F_{0o}^{-1}(u|p(X))\right) du\right]. \qquad (14)\end{aligned}$$

[13] We are grateful to an anonymous referee for pointing out the necessary and sufficient condition.
[14] A similar result can be established for φ-indicator functions. To save space, it is omitted from the paper.

If for every $x \in \mathcal{X}$, the conditional distribution functions of Y_1, Y_0 given $X = x$ are the same as the conditional distribution functions of Y_1, Y_0 given $p(X) = p(x)$, then the identified set for θ_o based on the propensity score is identical to the identified set based on X; otherwise the former is in general larger than the latter.

Proposition 4.1 *Suppose that Assumption (IX) holds. For a strict super-modular and right continuous function $\mu(\cdot,\cdot)$, suppose that the four expectations in (14) and (9) with $X^* = X$ exist (even if infinite valued) and that either of the conditions (a) and (b) in Theorem 3.1 (with θ_{LP} and θ_{UP} replacing θ^L and θ^U in condition (b)) is satisfied. Then $\theta_{LP} = \theta_L$ and $\theta_U = \theta_{UP}$ iff for μ_c-almost all (y_1, y_0), it holds that*

$$\begin{aligned} &\Pr\left(F_{1o}(y_1|X) + F_{0o}(y_0|X) - 1 > 0 | p(X)\right) \in \{0,1\} \text{ and} \\ &\Pr\left(F_{1o}(y_1|X) - F_{0o}(y_0|X) < 0 | p(X)\right) \in \{0,1\}. \end{aligned} \tag{15}$$

Proposition 4.1 shows that for parameter θ_o defined by a strict super-modular function, the use of the full vector of covariates X shrinks the identified set using the propensity score $p(X)$ unless the conditional distributions $F_{1o}(y_1|X), F_{0o}(y_0|X)$ satisfy (15) which holds if the conditional marginal cdfs of Y_1, Y_0 depend on X only through $p(X)$.[15]

In the latent threshold-crossing model (3), $X^* = (X', \epsilon)'$ and the lower or upper bounds in Theorem 3.2 (i) are reached when the two potential outcomes are perfectly negatively or positively dependent conditional on X^*. For example, if $\epsilon = F(U_1 - U_0)$ (where F is cdf of $U_1 - U_0$) and $g_1(\cdot,\cdot)$, $g_0(\cdot,\cdot)$ are increasing (or decreasing) respectively in U_1 and U_0, then Y_1, Y_0 are perfectly positively dependent conditional on X^* and the upper bound is reached. When the distribution of either Y_1 or Y_0 conditional on X^* is degenerate, the lower and upper bounds in Theorem 3.2 (i) coincide and thus point identify θ_o. The following proposition follows from a similar proof to that of Theorem 3.3 or Proposition 4.1.

Proposition 4.2 *Suppose that Assumptions (IU) and (LS) hold. For a strict super-modular and right continuous function $\mu(\cdot,\cdot)$, suppose that the four expectations in (16) exist (even if infinite valued) and that either of the conditions (a) and (b) in Theorem 3.1 (with $E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X), F_{0o}^{-1}(1-u|X)\right) du\right]$ and $E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X), F_{0o}^{-1}(u|X)\right) du\right]$ replacing θ^L and θ^U in condition (b)) is satisfied. Then*

$$\begin{aligned} &E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X^*), F_{0o}^{-1}(1-u|X^*)\right) du\right] = E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X), F_{0o}^{-1}(1-u|X)\right) du\right] \text{ and} \\ &E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X^*), F_{0o}^{-1}(u|X^*)\right) du\right] = E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X), F_{0o}^{-1}(u|X)\right) du\right] \end{aligned} \tag{16}$$

iff for μ_c-almost all (y_1, y_0), it holds that

$$\begin{aligned} &\Pr\left(F_{1o}(y_1|X^*) + F_{0o}(y_0|X^*) - 1 > 0 | X\right) \in \{0,1\} \text{ and} \\ &\Pr\left(F_{1o}(y_1|X^*) - F_{0o}(y_0|X^*) < 0 | X\right) \in \{0,1\}. \end{aligned} \tag{17}$$

[15] For the point identified ATE, it is known that matching on the propensity score may result in loss of efficiency, see Hahn (1998, 2004).

Proposition 4.2 implies that in general taking into account the self-selection process in addition to the covariate X in the latent threshold-crossing model is more informative than using X only unless (17) holds**.** For instance, if U_1 and U_0 are independent of ϵ given X, Z, implying that both the conditional cdfs of Y_1, Y_0 given X^* are the same as those given X, then (17) holds. Note that in the latent threshold-crossing model, the distribution of ϵ conditional on X is $U(0,1)$. Thus both expectations with respect X^* in (16) can be expressed as follows:

$$E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X^*), F_{0o}^{-1}(1-u|X^*)\right)du\right] = E\left[\int_0^1\int_0^1 \mu\left(F_{1o}^{-1}(u|X,v), F_{0o}^{-1}(1-u|X,v)\right)dudv\right]$$
$$E\left[\int_0^1 \mu\left(F_{1o}^{-1}(u|X^*), F_{0o}^{-1}(u|X^*)\right)du\right] = E\left[\int_0^1\int_0^1 \mu\left(F_{1o}^{-1}(u|X,v), F_{0o}^{-1}(u|X,v)\right)dudv\right].$$

We now provide a detailed analysis of the identified set for the correlation coefficient in a latent threshold-crossing model to demonstrate the role of endogenous selection in shrinking the identified set. When there is one observable covariate X, Example (i)-(IC) in Section 3 establishes the condition: $\rho_{0X}^2 + \rho_{1X}^2 > 1$ under which the sign of the correlation coefficient is identified. We show in Example (i)-(IU) below that this condition may be weakened in a specific latent threshold-crossing model with endogenous selection.[16]

Example (i)-(IU) (Correlation Coefficient). Consider the following special case of the latent threshold-crossing model (3):

$$Y_1 = g_1(X) + U_1,\ Y_0 = g_0(X) + U_0,\ \text{and } D = I\{g(Z) - \epsilon > 0\}.$$

Since the distribution of ϵ conditional on X is normalized to be $U(0,1)$, the distribution of $V \equiv \Phi^{-1}(\epsilon)$ conditional on X is $N(0,1)$, where $\Phi(\cdot)$ is the cdf of $N(0,1)$. Suppose that $(U_1, U_0, \epsilon)'$ is independent of Z conditional on X, implying that Assumptions (IU)-(ii) holds. Then the joint distribution of $(U_1,U_0,V,X,Z)'$ can be expressed as $f(u_1,u_0,v,x,z) = f(u_1,u_0,v,x)f(z|x)$. Thus we only need to consider the joint distribution of $(U_1,U_0,V,X)'$. Let $U = (U_1,U_0)'$, $X^* = (V,X)'$ and assume for simplicity that $g_i(X) = \mu_i$ $(i = 1,0)$ are constants and $(U_1,U_0,V,X)'$ follows a multivariate normal distribution:

$$\begin{pmatrix} U \\ X^* \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]. \tag{18}$$

In Appendix C, we demonstrate that (1) when there is endogenous selection, the identified set for ρ_{10}, denoted by $\Theta_{IC}^* = \left[\rho_L^{(2)}, \rho_U^{(2)}\right]$ given in (C.9) and (C.10), is smaller than that without endogenous selection as in Example (i)-(IC), denoted by $\Theta_{IC} = \left[\rho_L^{(1)}, \rho_U^{(1)}\right]$ shown in (C.7) and (C.8); (2) as long as the correlations between $V \equiv \Phi^{-1}(\epsilon)$ and U_j (i.e., ρ_{1V} and ρ_{0V}) are strong enough so that $\rho_{0V}^2 + \rho_{1V}^2 > 1$, the bounds $\rho_L^{(2)}$ and $\rho_U^{(2)}$ with endogenous selection are able to identify the sign of ρ_{10} under quite weak conditions on the dependence between (Y_1, Y_0) and the observable covariate X: $\rho_{10} > 0$ when $\rho_{0X}\rho_{1X} \geq 0$ with $\rho_{1V}\rho_{0V} > 0$, and $\rho_{10} < 0$ when $\rho_{0X}\rho_{1X} \leq 0$ with $\rho_{1V}\rho_{0V} < 0$; (3) Example (i)-(IU) also validates Proposition 4.2.

[16] The on-line Supplementary Appendices offer another such example.

5 Concluding Remarks

In this paper, we have provided a comprehensive study of partial identification of $\theta_o \equiv E_o[\mu(Y_1, Y_0)]$ for two general classes of functions μ when only partial information on the joint distribution of (Y_1, Y_0) is available to the econometrician, see Assumption (IC). We have shown that the two commonly used frameworks to identify average treatment effects in the literature, i.e., the selection-on-observables and latent threshold-crossing models, satisfy Assumption (IC). The main contributions of this paper include: i) we establish the identified sets for functionals in both classes under various maintained assumptions and characterize conditions under which our identified sets point identify the true value of the functionals; ii) we establish sufficient and necessary conditions for the covariate information to tighten the identified sets without the covariate information; and iii) we characterize the role of the propensity score in the selection-on-observables framework and the role of endogenous selection in the latent threshold-crossing model.

Empirical applications of the results in this paper abound in economics, finance, and actuarial mathematics. In the context of evaluating treatment effects using latent threshold-crossing model, the results in this paper allow us to go beyond the analysis in Heckman and Vytlacil (2005), Carneiro and Lee (2009). Consider the labor market setting studied in Vijverberg (1993) in which the two treatment states are two different labor market sectors and Y_j is the wage offer in sector j, $j = 1, 0$. Assume Y_j is an accurate measure of productivity. Then the analogs of all the quantities discussed in Vijverberg (1993) for the Gaussian Switching Regime Model can be bounded using the results in this paper for the latent threshold-crossing model. Examples include: (i) out of the workers who would be more productive in sector 1, i.e., for whom $Y_1 > Y_0$, the share that is actually employed in sector 1; (ii) the distribution of the potential outcome Y_1 (productivity in sector 1) of an individual with an above average Y_0 (productivity in sector 0); and (iii) the distribution of the potential outcome Y_1 of an individual with an above average Y_0 who selects into the program. We refer interested readers to Vijverberg (1993) for more examples.

Extensions of the results in this paper include identification analysis for the same classes of functions when the sampling scheme only partially identifies the conditional marginal distribution of each outcome variable and the development of valid inference procedures for the distributional treatment effect parameters in latent threshold-crossing models. The authors are currently working on these.

Appendix A: Technical Proofs for Sections 3 and 4

Proof of Theorem 3.1: Noting that

$$\begin{aligned}\theta^U &= \int_0^{1/2} \mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) du + \int_{1/2}^{1} \mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) du \\ &= \int_0^{1/2} \mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) du + \int_0^{1/2} \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(1-u)\right) du\end{aligned}$$

and

$$\theta^L = \int_0^{1/2} \mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(1-u)\right) du + \int_0^{1/2} \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(u)\right) du,$$

we obtain

$$\theta^U - \theta^L = \int_0^{1/2} \left[\begin{array}{c} \mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) + \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(1-u)\right) \\ -\mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(1-u)\right) - \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(u)\right) \end{array} \right] du. \tag{A.1}$$

(i) If $\mu(\cdot,\cdot)$ is additively separable in its arguments, then $\theta^L = \theta^U$ follows directly from additive separability of $\mu(\cdot,\cdot)$. **(ii)** If $\mu(\cdot,\cdot)$ is super-modular, we have

$$\begin{aligned}&\mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) + \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(1-u)\right) \\ &-\mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(1-u)\right) - \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(u)\right) \geq 0, \ \forall u \in [0,1]\,.\end{aligned} \tag{A.2}$$

It then follows from (A.1) and (A.2) that $\theta^L = \theta^U$ if and only if for almost all $u \in [0, 1/2]$,

$$\begin{aligned}&\mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) + \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(1-u)\right) \\ &-\mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(1-u)\right) - \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(u)\right) = 0\,.\end{aligned} \tag{A.3}$$

Obviously, (A.3) holds when one of $F_{1o}(\cdot)$ and $F_{0o}(\cdot)$ is degenerate. Now we show $\theta^L = \theta^U$ implies that at least one of $F_{1o}(\cdot)$ and $F_{0o}(\cdot)$ is degenerate. Suppose that both $F_{1o}(\cdot)$ and $F_{0o}(\cdot)$ are non-degenerate. Then there are y_j, y_j' $(j = 1, 0)$ satisfying: $y_j < y_j'$ and $0 < F_{jo}(y_j) \leq F_{jo}\left(y_j'\right) < 1$. Define

$$u_* \equiv \min\left\{F_{1o}(y_1), 1 - F_{1o}(y_1'), F_{0o}(y_0), 1 - F_{0o}(y_0')\right\}.$$

Then $0 < u_* \leq 1/2$, and for all $u \in [0, u_*)$, we have

$$F_{jo}^{-1}(u) \leq y_j < y_j' < F_{jo}^{-1}(1-u)\,.$$

It follows from the "strict super-modular" assumption that

$$\begin{aligned}&\mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(u)\right) + \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(1-u)\right) \\ &-\mu\left(F_{1o}^{-1}(u), F_{0o}^{-1}(1-u)\right) - \mu\left(F_{1o}^{-1}(1-u), F_{0o}^{-1}(u)\right) > 0, \ \forall u \in [0, u_*)\,.\end{aligned}$$

This contradicts with (A.3), a sufficient and necessary condition for $\theta^L = \theta^U$ to hold. ■

Proof of Theorem 3.2: **(i)** For $x^* \in \mathcal{X}^*$, let $\theta_o(x^*) = E_o\left[\mu(Y_1, Y_0) | X^* = x^*\right]$,

$$\theta_L(x^*) = \int_0^1 \mu\left(F_{1o}^{-1}(u|x^*), F_{0o}^{-1}(1-u|x^*)\right) du = \int\int \mu(y_1, y_0)\, dM\left(F_{1o}(y_1|x^*), F_{0o}(y_0|x^*)\right),$$

and

$$\theta_U(x^*) = \int_0^1 \mu\left(F_{1o}^{-1}(u|x^*), F_{0o}^{-1}(u|x^*)\right) du = \int\int \mu(y_1, y_0)\, dW\left(F_{1o}(y_1|x^*), F_{0o}(y_0|x^*)\right).$$

Then $\theta_L = E[\theta_L(X^*)]$ and $\theta_U = E[\theta_U(X^*)]$. It follows from conditions in (A) and (B) that for almost all $x^* \in \mathcal{X}^*$, in case (A) $E[\mu(Y_1, Y_1)|X^* = x^*]$ and $E[\mu(Y_0, Y_0)|X^* = x^*]$ are finite, and in case (B), $E[\mu(Y_1, \overline{y}_0)|X^* = x^*]$, $E[\mu(\overline{y}_1, Y_0)|X^* = x^*]$, and at least one of $\theta_L(x^*)$ and $\theta_U(x^*)$ are finite. Thus, for both cases (A) and (B), Theorem 2 in Cambanis, Simons, and Stout (abbreviated to CSS) (1976) implies: $\theta_L(x^*) \leq \theta_o(x^*) \leq \theta_U(x^*)$ for all $x^* \in \mathcal{X}^*$. Taking expectations with respect to X^* leads to $\Theta_{IC} \subseteq [E(\theta_L(X^*)), E(\theta_U(X^*))] = [\theta_L, \theta_U]$.

Now we show that $[E(\theta_L(X^*)), E(\theta_U(X^*))] \subseteq \Theta_{IC}$, that is, for any given $V \in [E(\theta_L(X^*)), E(\theta_U(X^*))]$, there exists a conditional distribution $F_V(y_1, y_0|x^*) \equiv C_V(F_{1o}(y_1|x^*), F_{0o}(y_0|x^*)|x^*)$ with marginals $F_{1o}(y_1|x^*)$ and $F_{0o}(y_0|x^*)$ such that $E\left[\int\int \mu(y_1, y_0)\, dF_V(y_1, y_0|X^*)\right] = V$. Obviously, if $V = E(\theta_L(X^*))$ or $E(\theta_U(X^*))$, we take $C_V = M$ or W. Therefore, without loss of generality, suppose $E(\theta_L(X^*)) < V < E(\theta_U(X^*))$. If both $E(\theta_L(X^*))$ and $E(\theta_U(X^*))$ are finite, implying for almost all $x^* \in \mathcal{X}^*$ that both $\theta_L(x^*)$ and $\theta_U(x^*)$ are finite, then we can define

$$v = \frac{V - E(\theta_L(X^*))}{E(\theta_U(X^*)) - E(\theta_L(X^*))} \in (0, 1),$$

and

$$F_V(y_1, y_0|x^*) = vW\left(F_{1o}(y_1|x^*), F_{0o}(y_0|x^*)\right) + (1 - v)\, M\left(F_{1o}(y_1|x^*), F_{0o}(y_0|x^*)\right).$$

Obviously, $F_V(y_1, y_0|x^*)$ is a joint cdf conditional on $X^* = x^*$ with marginals $F_{1o}(y_1|x^*)$ and $F_{0o}(y_0|x^*)$, and satisfies

$$\int\int \mu(y_1, y_0)\, dF_V(y_1, y_0|x^*) = v\theta_U(x^*) + (1 - v)\,\theta_L(x^*)$$

and thus

$$E\left[\int\int \mu(y_1, y_0)\, dF_V(y_1, y_0|X^*)\right] = vE[\theta_U(X^*)] + (1 - v)\, E[\theta_L(X^*)] = V.$$

Now we consider the case that either of $E(\theta_L(X^*))$ and $E(\theta_U(X^*))$ is infinite, say, $-\infty = E(\theta_L(X^*)) < E(\theta_U(X^*)) < +\infty$. Notice that in case (A) we definitely have $\theta_U = E(\theta_U(X^*)) < +\infty$. To see this, from equation (5) in CSS (1976), we have

$$2\theta_U(X^*) = E[\mu(Y_1, Y_1)|X^*] + E[\mu(Y_0, Y_0)|X^*] - \int\int A_W^* d\mu_c(y_1, y_0),$$

where

$$\begin{aligned} A_W^* \equiv\ & F_{1o}(y_1 \wedge y_0|X^*) + F_{0o}(y_1 \wedge y_0|X^*) \\ & - W\left(F_{1o}(y_1 \vee y_0|X^*), F_{0o}(y_1 \wedge y_0|X^*)\right) \\ & - W\left(F_{1o}(y_1 \wedge y_0|X^*), F_{0o}(y_1 \vee y_0|X^*)\right). \end{aligned}$$

Taking expectations with respect to X^* leads to

$$2\theta_U = E[\mu(Y_1, Y_1)] + E[\mu(Y_0, Y_0)] - \int\int E[A_W^*]\, d\mu_c(y_1, y_0), \tag{A.4}$$

implying $\theta_U < +\infty$ in case (A) because $E[\mu(Y_1, Y_1)]$ and $E[\mu(Y_0, Y_0)]$ are finite and $A^*_W \geq 0$ for all y_1, y_0 and $X^* = x^*$. Now we show that there exists a conditional joint distribution $F_\alpha(y_1, y_0|x^*)$ with marginals $F_{1o}(y_1|x^*)$ and $F_{0o}(y_0|x^*)$ such that $-\infty < E(\theta_\alpha(X^*)) < V < E(\theta_U(X^*)) < +\infty$, where $E(\theta_\alpha(X^*)) \equiv E\left[\int\int \mu(y_1, y_0)\, dF_\alpha(y_1, y_0|X^*)\right]$. Actually, from the proof of Lemma in CSS (1976), for each $\alpha \in (0, 1/2]$, we can define

$$g_\alpha(u|x^*) = \begin{cases} F^{-1}_{0o}(1-u|x^*), & \text{if } \alpha \leq u \leq 1-\alpha, \\ F^{-1}_{0o}(u|x^*), & \text{if } 0 \leq u < \alpha \text{ or } 1-\alpha < u \leq 1, \end{cases}$$

and let $F_\alpha(\cdot, \cdot|x^*)$ be the joint distribution of $\left[F^{-1}_{1o}(U|x^*), g_\alpha(U|x^*)\right]$, where U is a uniform r.v. on $(0, 1)$. It is easy to show that $F_\alpha(y_1, y_0|x^*)$ has marginals $F_{1o}(y_1|x^*)$ and $F_{0o}(y_0|x^*)$, and that

$$\begin{aligned} E(\theta_\alpha(X^*)) = {} & E\left[\int_\alpha^{1-\alpha} \mu\left(F^{-1}_{1o}(u|X^*), F^{-1}_{0o}(1-u|X^*)\right) du\right] \\ & + E\left[\left(\int_0^\alpha + \int_{1-\alpha}^1\right) \mu\left(F^{-1}_{1o}(u|X^*), F^{-1}_{0o}(u|X^*)\right) du\right]. \end{aligned}$$

Note that the first part changes from zero to $E(\theta_L(X^*)) = -\infty$ as α decreases from $1/2$ to zero, but the second part is always finite and goes to zero. Thus, there exists an $\alpha \in (0, 1/2]$ such that $-\infty < E(\theta_\alpha(X^*)) < V < E(\theta_U(X^*))$. Similar to the argument above, we can define

$$F_V(y_1, y_0|x^*) = vW\left(F_{1o}(y_1|x^*), F_{0o}(y_0|x^*)\right) + (1-v) F_\alpha(y_1, y_0|X^*)$$

with $v = [V - E(\theta_\alpha(X^*))]/[E(\theta_U(X^*)) - E(\theta_\alpha(X^*))]$. These bounds are sharp, as they are achieved at $M(F_{1o}(\cdot|x^*), F_{0o}(\cdot|x^*)), W(F_{1o}(\cdot|x^*), F_{0o}(\cdot|x^*))$ respectively.

(ii) Define

$$\begin{aligned} \Delta(u|x^*) \equiv {} & \mu\left(F^{-1}_{1o}(u|x^*), F^{-1}_{0o}(u|x^*)\right) + \mu\left(F^{-1}_{1o}(1-u|x^*), F^{-1}_{0o}(1-u|x^*)\right) \\ & - \mu\left(F^{-1}_{1o}(u|x^*), F^{-1}_{0o}(1-u|x^*)\right) - \mu\left(F^{-1}_{1o}(1-u|x^*), F^{-1}_{0o}(u|x^*)\right). \end{aligned}$$

Similar to (A.2) and (A.1), we have $\Delta(u|x^*) \geq 0$ for all u and x^*, and

$$\theta_U - \theta_L = E\left[\int_0^{1/2} \Delta(u|X^*)\, du\right] = \int\left(\int_0^{1/2} \Delta(u|x^*)\, du\right) dF_{X^*o}(x^*) \geq 0.$$

Obviously, $\theta_U = \theta_L$ if and only if $\Delta(u|X^*) = 0$ with probability one for almost all $u \in [0, 1/2]$. When one of $F_{1o}(\cdot|x^*)$ and $F_{0o}(\cdot|x^*)$ is degenerate for almost all $x^* \in \mathcal{X}^*$, we have $\Delta(u|x^*) = 0$ for almost all u and x^*, implying $\theta_U = \theta_L$. Now we show under the "strict super-modular" assumption that if $\theta_U = \theta_L$, then one of $F_{1o}(\cdot|x^*)$ and $F_{0o}(\cdot|x^*)$ is degenerate for almost all $x^* \in \mathcal{X}^*$. By contradiction, assuming that there is a set $A \subset \mathcal{X}^*$ such that $\Pr(A) > 0$ and for every $x^* \in A$ both $F_{1o}(\cdot|x^*)$ and $F_{0o}(\cdot|x^*)$ are non-degenerate, then by a similar proof to that of Theorem 3.1 (ii), we have $\int_0^{1/2} \Delta(u|x^*)\, du > 0$ for every $x^* \in A$, implying $\theta_U - \theta_L > 0$, which is a contradiction. ■

Proof of Theorem 3.3: First, we show $\Theta_{IC} = [\theta_L, \theta_U] \subseteq \Theta_I = \left[\theta^L, \theta^U\right]$. Recall definitions of $F^{(-)}(y_1, y_0)$, $F^{(+)}(y_1, y_0)$, $F^{(-)}_*(y_1, y_0)$ and $F^{(+)}_*(y_1, y_0)$ in Subsection 3.1. For every (y_1, y_0), by Jensen's

inequality, we have

$$\begin{aligned} F^{(-)}(y_1, y_0) &= \max\{E[F_{1o}(y_1|X^*) + F_{0o}(y_0|X^*) - 1], 0\} \\ &\leq E[\max\{F_{1o}(y_1|X^*) + F_{0o}(y_0|X^*) - 1, 0\}] \\ &= F_*^{(-)}(y_1, y_0), \end{aligned} \tag{A.5}$$

and

$$\begin{aligned} F^{(+)}(y_1, y_0) &= E[F_{0o}(y_0|X^*)] + \min\{E[F_{1o}(y_1|X^*) - F_{0o}(y_0|X^*)], 0\} \\ &\geq E[F_{0o}(y_0|X^*) + \min\{F_{1o}(y_1|X^*) - F_{0o}(y_0|X^*), 0\}] \\ &= F_*^{(+)}(y_1, y_0). \end{aligned} \tag{A.6}$$

Under condition (a) of Theorem 3.1, it follows from equation (5) in CSS (1976) that we have

$$2\theta^U = E[\mu(Y_1, Y_1)] + E[\mu(Y_0, Y_0)] - \iint A_W d\mu_c(y_1, y_0), \tag{A.7}$$

where

$$\begin{aligned} A_W = {} & F_{1o}(y_1 \wedge y_0) + F_{0o}(y_1 \wedge y_0) \\ & - F^{(+)}(y_1 \vee y_0, y_1 \wedge y_0) - F^{(+)}(y_1 \wedge y_0, y_1 \vee y_0). \end{aligned}$$

Note that $E[A_W^*]$ in (A.4) can be expressed as

$$\begin{aligned} E[A_W^*] = {} & F_{1o}(y_1 \wedge y_0) + F_{0o}(y_1 \wedge y_0) \\ & - F_*^{(+)}(y_1 \vee y_0, y_1 \wedge y_0) - F_*^{(+)}(y_1 \wedge y_0, y_1 \vee y_0). \end{aligned}$$

By (A.6), we have $E[A_W^*] \geq A_W$ for all (y_1, y_0). Comparing (A.4) and (A.7) leads to $\theta_U \leq \theta^U$. Similarly, we can show $\theta^L \leq \theta_L$ by using the following results:

$$\begin{aligned} 2\theta^L &= E[\mu(Y_1, Y_1)] + E[\mu(Y_0, Y_0)] - \iint A_M d\mu_c(y_1, y_0), \\ 2\theta_L &= E[\mu(Y_1, Y_1)] + E[\mu(Y_0, Y_0)] - \iint E[A_M^*] d\mu_c(y_1, y_0), \end{aligned}$$

and $E[A_M^*] \leq A_M$ for all (y_1, y_0), where

$$\begin{aligned} A_M &= F_{1o}(y_1 \wedge y_0) + F_{0o}(y_1 \wedge y_0) - F^{(-)}(y_1 \vee y_0, y_1 \wedge y_0) - F^{(-)}(y_1 \wedge y_0, y_1 \vee y_0), \\ E[A_M^*] &= F_{1o}(y_1 \wedge y_0) + F_{0o}(y_1 \wedge y_0) - F_*^{(-)}(y_1 \vee y_0, y_1 \wedge y_0) - F_*^{(-)}(y_1 \wedge y_0, y_1 \vee y_0), \end{aligned}$$

and (A.5) is used. Combining $\theta^L \leq \theta_L$ and $\theta_U \leq \theta^U$ implies $\Theta_{IC} \subset \Theta_I$.

Under condition (b) of Theorem 3.1, it follows from equation (9) in CSS (1976) that

$$\theta^L = E[\mu(Y_1, \overline{y}_0)] + E[\mu(\overline{y}_1, Y_0)] - \mu(\overline{y}_1, \overline{y}_0) + \iint B_M d\mu_c(y_1, y_0) \text{ and} \tag{A.8}$$

$$\theta_L(X^*) = E[\mu(Y_1, \overline{y}_0)|X^*] + E[\mu(\overline{y}_1, Y_0)|X^*] - \mu(\overline{y}_1, \overline{y}_0) + \iint B_M^* d\mu_c(y_1, y_0), \tag{A.9}$$

where for all (y_1, y_0),

$$\begin{aligned}
B_M &= M(F_{1o}(y_1), F_{0o}(y_0)) - 1(\overline{y}_1 < y_1) F_{0o}(y_0) \\
&\quad - F_{1o}(y_1) 1(\overline{y}_0 < y_0) + 1(\overline{y}_1 < y_1) 1(\overline{y}_0 < y_0), \\
B_M^* &= M(F_{1o}(y_1|X^*), F_{0o}(y_0|X^*)) - 1(\overline{y}_1 < y_1) F_{0o}(y_0|X^*) \\
&\quad - F_{1o}(y_1|X^*) 1(\overline{y}_0 < y_0) + 1(\overline{y}_1 < y_1) 1(\overline{y}_0 < y_0).
\end{aligned}$$

Taking expectations between both sides of (A.9) with respect to X^*, we also have

$$\theta_L = E[\mu(Y_1, \overline{y}_0)] + E[\mu(\overline{y}_1, Y_0)] - \mu(\overline{y}_1, \overline{y}_0) + \iint E(B_M^*) d\mu_c, \tag{A.10}$$

where $E(B_M^*) = F_*^{(-)}(y_1, y_0) - 1(\overline{y}_1 < y_1) F_{0o}(y_0) - F_{1o}(y_1) 1(\overline{y}_0 < y_0) + 1(\overline{y}_1 < y_1) 1(\overline{y}_0 < y_0)$ for all (y_1, y_0). Note from (A.5) that $E(B_M^*) \geq B_M$ for all (y_1, y_0). Then, by comparing (A.8) and (A.10), we have $\theta^L \leq \theta_L$. Similarly, for the upper bounds, we can show $\theta_U \leq \theta^U$. Both $\theta^L \leq \theta_L$ and $\theta_U \leq \theta^U$ imply $\Theta_{IC} \subset \Theta_I$.

Now we present sufficient and necessary conditions for $\Theta_{IC} = \Theta_I$. If $\mu(\cdot, \cdot)$ is strict super-modular (implying that any rectangle in (y_1, y_0)-plane has a positive μ_c measure), it follows from (A.5) and (A.6) that in both cases $\theta_L = \theta^L$ iff $F_*^{(-)}(y_1, y_0) = F^{(-)}(y_1, y_0)$ for μ_c-almost all (y_1, y_0) and $\theta_U = \theta^U$ iff $F_*^{(+)}(y_1, y_0) = F^{(+)}(y_1, y_0)$ for μ_c-almost all (y_1, y_0). Furthermore, for μ_c-almost every (y_1, y_0), $F_*^{(-)}(y_1, y_0) = F^{(-)}(y_1, y_0)$ iff $\Pr(F_{1o}(y_1|X^*) + F_{0o}(y_0|X^*) - 1 > 0) \in \{0, 1\}$ and $F_*^{(+)}(y_1, y_0) = F^{(+)}(y_1, y_0)$ iff $\Pr(F_{1o}(y_1|X^*) - F_{0o}(y_0|X^*) < 0) \in \{0, 1\}$. ■

Proof of Theorem 3.4: First, we introduce some notation from the literature, see Frank, Nelsen, and Schweizer (1987), Williamson and Downs (1990), and Embrechts, Hoeing, and Juri (2003). For any bivariate copula function C and any univariate cdfs F_1, F_0, let

$$\begin{aligned}
\sigma_{C,\varphi}(F_1, F_0)(\delta) &\equiv \int_{\{\varphi(u,v) < \delta\}} dC(F_1(u), F_0(v)), \\
\tau_{C,\varphi}(F_1, F_0)(\delta) &\equiv \sup_{\varphi(u,v)=\delta} C(F_1(u), F_0(v)), \text{ and} \\
\rho_{C,\varphi}(F_1, F_0)(\delta) &\equiv \inf_{\varphi(u,v)=\delta} C^d(F_1(u), F_0(v)),
\end{aligned}$$

where C^d is the dual of C, $C^d(x, y) \equiv x + y - C(x, y)$ for all $x, y \in [0, 1]$. Since Y_1, Y_0 are continuous random variables and φ is continuous, from Theorem 5.1 of Frank, Nelsen, and Schweizer (1987), we have

$$\sigma_{C,\varphi}(F_{1o}, F_{0o})(\delta) = F_\varphi(\delta),$$

$$\begin{aligned}
\tau_{M,\varphi}(F_{1o}, F_{0o})(\delta) &= \sup_{\varphi(u,v)=\delta} \max\{F_{1o}(u) + F_{0o}(v) - 1, 0\} \\
&= \sup_y \max\left\{F_{1o}(y) + F_{0o}(\hat{\varphi_y}(\delta)) - 1, 0\right\} = F_{\min,\varphi}(\delta),
\end{aligned}$$

$$\begin{aligned}
\rho_{M,\varphi}(F_{1o}, F_{0o})(\delta) &\equiv \inf_{\varphi(u,v)=\delta} [F_{1o}(u) + F_{0o}(v) - M(F_{1o}(u), F_{0o}(v))] \\
&= 1 + \inf_y \min\left\{F_{1o}(y) + F_{0o}(\hat{\varphi_y}(\delta)) - 1, 0\right\} = F_{\max,\varphi}(\delta),
\end{aligned}$$

and for any δ and any copula $C(\cdot,\cdot)$ or joint distribution $F(u,v) = C(F_{1o}(u), F_{0o}(v))$ with marginals F_{1o} and F_{0o},

$$F_{\min,\varphi}(\delta) = \tau_{M,\varphi}\left(F_{1o}, F_{0o}\right)(\delta) \leq F_{\varphi}\left(\delta\right) \leq \rho_{M,\varphi}\left(F_{1o}, F_{0o}\right)(\delta) = F_{\max,\varphi}(\delta), \tag{A.11}$$

where we used the assumption that Y_1 and Y_0 are continuous random variables, implying $\rho_{M,\varphi}\left(F_{1o}, F_{0o}\right)(\delta+) = \rho_{M,\varphi}\left(F_{1o}, F_{0o}\right)(\delta)$.

It is easy to verify that $F_{\min,\varphi}(\delta) = F_{\varphi}(\delta) = F_{\max,\varphi}(\delta)$ for all δ if either F_{1o} or F_{0o} is a degenerate distribution. By a straightforward extension of the argument used to establish Corollary 2 of Theorem 9 in Moynihan, Schweizer, and Sklar (abbreviated to MSS) (1978), it follows that $F_{\varphi}\left(\delta\right) = F_{\min,\varphi}(\delta)$ implies that at least one of the marginal distributions F_{1o}, F_{0o} is degenerate, and similarly this is true for $F_{\varphi}\left(\delta\right) = F_{\max,\varphi}(\delta)$. Here we only show that if neither F_{1o} nor F_{0o} is degenerate, then $F_{\varphi}\left(\delta\right) < F_{\max,\varphi}(\delta)$ for some δ. In fact, by Corollary of Theorem 3 in MSS (1978), we have

$$F_{\varphi} = \sigma_{C,\varphi}\left(F_{1o}, F_{0o}\right) \leq \rho_{C,\varphi}\left(F_{1o}, F_{0o}\right) \leq \rho_{M,\varphi}\left(F_{1o}, F_{0o}\right) = F_{\max,\varphi}. \tag{A.12}$$

If $C(F_{1o}(u), F_{0o}(v)) = W(F_{1o}(u), F_{0o}(v))$ for all (u,v), then by Theorem 6 and Corollary of Theorem 10 in MSS (1978), respectively, we have $\sigma_{C,\varphi}\left(F_{1o}, F_{0o}\right) = \rho_{W,\varphi}\left(F_{1o}, F_{0o}\right)$ and $\rho_{W,\varphi}\left(F_{1o}, F_{0o}\right) < \rho_{M,\varphi}\left(F_{1o}, F_{0o}\right)$ (because of $M(a,b) < W(a,b)$ for all (a,b) in $(0,1)\times(0,1)$), implying that the inequality in (A.12) is strict. If $C(F_{1o}(u), F_{0o}(v)) < W(F_{1o}(u), F_{0o}(v))$ for some (u,v), then by Theorem 8 in MSS (1978), we have $\sigma_{C,\varphi}\left(F_{1o}, F_{0}\right) < \rho_{C,\varphi}\left(F_{1o}, F_{0o}\right)$, also implying that the inequality in (A.12) is strict. ■

Proof of Theorem 3.5: First of all, we show that $F_{\min,\varphi}(\delta|X^*)$ and $F_{\max,\varphi}(\delta|X^*)$ are measurable for each δ. Since the supports of Y_1 and Y_0 given X^*, $\mathcal{Y}_1\left(X^*\right)$ and $\mathcal{Y}_0\left(X^*\right)$, are the Borel sets, without loss of generality, we can assume $\mathcal{Y}_1\left(X^*\right)$ and $\mathcal{Y}_0\left(X^*\right)$ are two intervals. Let $\mathcal{Y}_j\left(X^*\right) = \left(Q_{jL}\left(X^*\right), Q_{jR}\left(X^*\right)\right]$, $j = 1, 0$; and suppose that $Q_{jL}\left(X^*\right)$ and $Q_{jR}\left(X^*\right)$ are measurable. Obviously, $\hat{\varphi_y}\left(\delta|X^*\right)$ is also measurable for each δ. Let $\mathbb{Q}$ be the set of rationale numbers which is such that, for any real y, there is a sequence $\{y_k\} \subset \mathbb{Q}$ with $y \leq y_k \leq y + 1/k$ (take for instance $[ky+1]/k$ where $[\cdot]$ is the integer part). Since $P\left(\lim_{k\uparrow\infty} \mathbb{I}\left(Y \leq y_k\right) = \mathbb{I}\left(Y \leq y\right)\right) = 1$, the Lebesgue Dominated Convergence Theorem gives that $\lim_{k\uparrow\infty} F_{1o}(y_k|X^*) = F_{1o}(y|X^*)$, and since $\hat{\varphi_{y_{k-1/k}}}\left(\delta|X^*\right) \geq \hat{\varphi_y}\left(\delta|X^*\right)$ with $\lim_{k\uparrow\infty} \hat{\varphi_{y_{k-1/k}}}\left(\delta|X^*\right) = \hat{\varphi_y}\left(\delta|X^*\right)$, $\lim_{k\uparrow\infty} F_{0o}(\hat{\varphi_{y_{k-1/k}}}\left(\delta|X^*\right)|X^*) = F_{0o}(\hat{\varphi_y}\left(\delta|X^*\right)|X^*)$. It then follows that

$$\begin{aligned} F_{\min,\varphi}(\delta|X^*) &= \lim_{k\uparrow\infty} \sup_{y\in(Q_{1L}(X^*),Q_{1R}(X^*)]\cap\mathbb{Q}} \max\{F_{1o}(y|X^*) + F_{0o}(\hat{\varphi_{y-1/k}}\left(\delta|X^*\right)|X^*) - 1, 0\} \\ &= \lim_{k\uparrow\infty} \sup_{y\in\mathbb{Q}} \max\left\{\left(F_{1o}(y|X^*) + F_{0o}(\hat{\varphi_{y-1/k}}\left(\delta|X^*\right)|X^*) - 1\right)\mathbb{I}\left[y \in \left(Q_{1L}\left(X^*\right), Q_{1R}\left(X^*\right)\right]\right], 0\right\}. \end{aligned}$$

This implies that $X^* \mapsto F_{\min,\varphi}(\delta|X^*)$ is measurable for any given δ, as obtained by taking limit and supremum of a countable number of measurable functions. That $X^* \mapsto F_{\max,\varphi}(\delta|X^*)$ is measurable similarly follows.

(i) It follows from (A.11) that $F_{\min,\varphi}(\delta|X^*) \leq F_{\varphi}(\delta|X^*) \leq F_{\max,\varphi}(\delta|X^*)$, and taking expectation with respect to X^* yields

$$E\left[F_{\min,\varphi}(\delta|X^*)\right] \leq F_{\varphi}\left(\delta\right) = E(F_{\varphi}(\delta|X^*)) \leq E\left[F_{\max,\varphi}(\delta|X^*)\right],$$

implying that $\Theta_{IC} \subseteq [E[F_{\min,\varphi}(\delta|X^*)], E[F_{\max,\varphi}(\delta|X^*)]]$. Now we show $[E[F_{\min,\varphi}(\delta|X^*)], E[F_{\max,\varphi}(\delta|X^*)]] \subseteq \Theta_{IC}$. Without loss of generality, suppose $E[F_{\min,\varphi}(\delta|X^*)] < E[F_{\max,\varphi}(\delta|X^*)]$, and for any given $V \in [E[F_{\min,\varphi}(\delta|X^*)], E[F_{\max,\varphi}(\delta|X^*)]]$, let

$$\alpha = \frac{E[F_{\max,\varphi}(\delta|X^*)] - V}{E[F_{\max,\varphi}(\delta|X^*)] - E[F_{\min,\varphi}(\delta|X^*)]} \in [0,1].$$

By Theorem 3 of Williamson and Downs (1990), there are copulas $C^{(t)}(u,v)$ and $C^{(r)}(u,v)$, depending only on the values of $t = F_{\min,\varphi}(\delta|x^*)$ and $r = F_{\max,\varphi}(\delta|x^*)$ respectively, such that

$$F_{\min,\varphi}(\delta|x^*) = \int_{\{\varphi(u,v)<\delta\}} dC^{(t)}(F_{1o}(u|x^*), F_{0o}(v|x^*)) \text{ and}$$
$$F_{\max,\varphi}(\delta|x^*) = \int_{\{\varphi(u,v)<\delta\}} dC^{(r)}(F_{1o}(u|x^*), F_{0o}(v|x^*)).$$

Define $F_V(u,v|x^*) = \alpha C^{(t)}(F_{1o}(u|x^*), F_{0o}(v|x^*)) + (1-\alpha)C^{(r)}(F_{1o}(u|x^*), F_{0o}(v|x^*))$, which is a joint distribution conditional on $X^* = x^*$ with marginals $F_{1o}(y_1|x^*)$ and $F_{0o}(y_0|x^*)$. Then

$$E_{F_V}[\mu(Y_1,Y_0)|X^* = x^*] = \int_{\{\varphi(u,v)<\delta\}} dF_V(u,v|x^*) = \alpha F_{\min,\varphi}(\delta|x^*) + (1-\alpha)F_{\max,\varphi}(\delta|x^*)$$

and thus

$$E(E_{F_V}[\mu(Y_1,Y_0)|X^*]) = \alpha E[F_{\min,\varphi}(\delta|X^*)] + (1-\alpha)E[F_{\max,\varphi}(\delta|X^*)] = V,$$

implying $V \in \Theta_{IC}$.

(ii) Obviously, the sufficient condition holds. Here we show the necessary condition, that is, when $E[F_{\max,\varphi}(\delta|X^*)] = E[F_{\min,\varphi}(\delta|X^*)]$, at least one of the conditional marginal distributions $F_{1o}(\cdot|x^*), F_{0o}(\cdot|x^*)$ is degenerate for almost every $x^* \in \mathcal{X}^*$. If not, then there exists a set $A \subset \mathcal{X}^*$ such that $\Pr(A) > 0$ and for all $x^* \in A$ both $F_{1o}(\cdot|x^*)$ and $F_{0o}(\cdot|x^*)$ are not degenerate, implying by Theorem 3.4 that we have $F_{\max,\varphi}(\delta|x^*) > F_{\min,\varphi}(\delta|x^*)$ for all $x^* \in A$, where we used the fact $F_{\max,\varphi}(\delta|x^*) \geq F_{\min,\varphi}(\delta|x^*)$ for all $x^* \in \mathcal{X}^*$. This leads to $E[F_{\max,\varphi}(\delta|X^*) - F_{\min,\varphi}(\delta|X^*)] > 0$, a contradiction with $E[F_{\max,\varphi}(\delta|X^*)] = E[F_{\min,\varphi}(\delta|X^*)]$. ■

Proof of Theorem 3.6. We provide a proof for the lower bounds. The proof for the upper bounds is similar and thus omitted. By definitions of $F_{L,\varphi}(\delta)$, $F_{\min,\varphi}(\delta)$ and Jensen's inequality, we obtain:

$$\begin{aligned}
F_{L,\varphi}(\delta) &= E\left[\sup_{y\in\mathcal{Y}_1} \max\left\{F_{1o}(y|X^*) + F_{0o}(\hat{\varphi_y}(\delta)|X^*) - 1, 0\right\}\right] \\
&\geq \sup_{y\in\mathcal{Y}_1} E\left[\max\left\{F_{1o}(y|X^*) + F_{0o}(\hat{\varphi_y}(\delta)|X^*) - 1, 0\right\}\right] \\
&\geq \sup_{y\in\mathcal{Y}_1} \max\left\{E\left[F_{1o}(y|X^*) + F_{0o}(\hat{\varphi_y}(\delta)|X^*) - 1\right], 0\right\} \\
&= \sup_{y\in\mathcal{Y}_1} \max\left\{F_{1o}(y) + F_{0o}(\hat{\varphi_y}(\delta)) - 1, 0\right\} = F_{\min,\varphi}(\delta).
\end{aligned}$$

Note that Y_j $(j = 1, 0)$ are assumed to be continuous random variables. Then, $F_{jo}(y|X^*)$ and $F_{jo}(y)$ are continuous and thus $\sup_{y\in\mathcal{Y}_1} F_{1o}(y|X^*) = 1$ and $\sup_{y\in\mathcal{Y}_1} F_{1o}(y) = 1$, implying $\sup_{y\in\mathcal{Y}_1}\{F_{1o}(y|X^*) +$

$F_{0o}(\hat{\varphi_y}(\delta)|X^*)-1\} \geq 0$ and $\sup_{y\in\mathcal{Y}_1}\{F_{1o}(y)+F_{0o}(\hat{\varphi_y}(\delta))-1\} \geq 0$. Therefore, the inequality above becomes

$$\begin{aligned} F_{L,\varphi}(\delta) &= E\left[\sup_{y\in\mathcal{Y}_1}\left\{F_{1o}(y|X^*)+F_{0o}(\hat{\varphi_y}(\delta)|X^*)-1\right\}\right] \\ &\geq \sup_{y\in\mathcal{Y}_1} E\left[F_{1o}(y|X^*)+F_{0o}(\hat{\varphi_y}(\delta)|X^*)-1\right] \\ &= \sup_{y\in\mathcal{Y}_1}\left[F_{1o}(y)+F_{0o}(\hat{\varphi_y}(\delta))-1\right] = F_{\min,\varphi}(\delta). \end{aligned} \tag{A.13}$$

Let $G_\varphi(y,x) = F_{1o}(y|x)+F_{0o}(\hat{\varphi_y}(\delta)|x)-1$. Then, $\sup_{y\in\mathcal{Y}_1} E[G(y,X^*)] = E[G(\overline{y},X^*)]$ and it follows from (A.13) that $F_{L,\varphi}(\delta) = F_{\min,\varphi}(\delta)$ iff $E[\sup_{y\in\mathcal{Y}_1} G(y,X^*)] = E[G(\overline{y},X^*)]$. Since $\sup_{y\in\mathcal{Y}_1} G(y,x^*) \geq G(\overline{y},x^*)$ for all $x^* \in \mathcal{X}^*$, it implies that $E\left[\sup_{y\in\mathcal{Y}_1} G(y,X^*)\right] = E[G(\overline{y},X^*)]$ iff $\sup_{y\in\mathcal{Y}_1} G(y,x^*) = G(\overline{y},x^*)$ for almost all $x^* \in \mathcal{X}^*$, that is, $F_{1o}(y|x^*)+F_{0o}(\hat{\varphi_y}(\delta)|x^*)-1$ reaches its maximum value uniformly at $\overline{y}$ for almost all $x^* \in \mathcal{X}^*$. ■

Proof of Proposition 4.1: By using Theorem 3.2 (i) with $X^* = p(X)$, we see that the identified set for θ_o is $[\theta_{LP},\theta_{UP}]$. Similarly, by using Theorem 3.2 (i) with $X^* = X$, we obtain the other identified set for θ_o, i.e., $[\theta_L,\theta_U]$. Following the proof of Theorem 3.3, we can show $[\theta_L,\theta_U] \subseteq [\theta_{LP},\theta_{UP}]$. Denote

$$\begin{aligned} F^L(y_1,y_0) &= E\left[M\left(F_{1o}(y_1|X),F_{0o}(y_0|X)\right)\right], \\ F_P^L(y_1,y_0) &= E\left[M\left(F_{1o}(y_1|p(X)),F_{0o}(y_0|p(X))\right)\right], \end{aligned}$$

$$\begin{aligned} F^U(y_1,y_0) &= E\left[W\left(F_{1o}(y_1|X),F_{0o}(y_0|X)\right)\right], \text{ and} \\ F_P^U(y_1,y_0) &= E\left[W\left(F_{1o}(y_1|p(X)),F_{0o}(y_0|p(X))\right)\right]. \end{aligned}$$

It follows from Jensen's inequality that for all (y_1,y_0), we have

$$\begin{aligned} F_P^L(y_1,y_0) &= E\left[\max\left\{F_1(y_1|p(X))+F_0(y_0|p(X))-1,0\right\}\right] \\ &= E\left[\max\left\{E\left[F_1(y_1|X)+F_0(y_0|X)-1|p(X)\right],0\right\}\right] \\ &\leq E\left[E\left[\max\left\{F_1(y_1|X)+F_0(y_0|X)-1,0\right\}|p(X)\right]\right] \\ &= E\left[\max\left\{F_1(y_1|X)+F_0(y_0|X)-1,0\right\}\right] \\ &= F^L(y_1,y_0), \end{aligned} \tag{A.14}$$

and

$$\begin{aligned} F_P^U(y_1,y_0) &= E\left[F_{0o}(y_0|p(X))+\min\left(F_{1o}(y_1|p(X))-F_{0o}(y_0|p(X)),0\right)\right] \\ &= E\left[E\left(F_{0o}(y_0|X)|p(X)\right)+\min\left(E\left(F_{1o}(y_1|X)-F_{0o}(y_0|X)|p(X)\right),0\right)\right] \\ &\geq E\left[E\left(F_{0o}(y_0|X)|p(X)\right)+E\left(\min\left(F_{1o}(y_1|X)-F_{0o}(y_0|X),0\right)|p(X)\right)\right] \\ &= E\left[F_{0o}(y_0|X)+\min\left(F_{1o}(y_1|X)-F_{0o}(y_0|X),0\right)\right] \\ &= F^U(y_1,y_0). \end{aligned}$$

To save space, we only consider condition (b) and show $\theta_{LP} \leq \theta_L$. Similar to (A.8) and (A.9), we have

$$\begin{aligned} \theta_{LP}(p(X)) &= E[\mu(Y_1, \overline{y}_0)|p(X)] + E[\mu(\overline{y}_1, Y_0)|p(X)] \qquad \text{(A.15)} \\ &\quad - \mu(\overline{y}_1, \overline{y}_0) + \iint BP_M d\mu_c(y_1, y_0) \text{ and} \\ \theta_L(X) &= E[\mu(Y_1, \overline{y}_0)|X] + E[\mu(\overline{y}_1, Y_0)|X] \qquad \text{(A.16)} \\ &\quad - \mu(\overline{y}_1, \overline{y}_0) + \iint BX_M d\mu_c(y_1, y_0), \end{aligned}$$

where for all (y_1, y_0),

$$\begin{aligned} BP_M &= M(F_{1o}(y_1|p(X)), F_{0o}(y_0|p(X))) - 1(\overline{y}_1 < y_1) F_{0o}(y_0|p(X)) \\ &\quad - F_{1o}(y_1|p(X)) 1(\overline{y}_0 < y_0) + 1(\overline{y}_1 < y_1) 1(\overline{y}_0 < y_0), \\ BX_M &= M(F_{1o}(y_1|X), F_{0o}(y_0|X)) - 1(\overline{y}_1 < y_1) F_{0o}(y_0|X) \\ &\quad - F_{1o}(y_1|X) 1(\overline{y}_0 < y_0) + 1(\overline{y}_1 < y_1) 1(\overline{y}_0 < y_0). \end{aligned}$$

Taking expectations for (A.15) and (A.16) with respect to X, we have

$$\begin{aligned} \theta_{LP} &= E[\theta_{LP}(p(X))] = E[\mu(Y_1, \overline{y}_0)] + E[\mu(\overline{y}_1, Y_0)] \qquad \text{(A.17)} \\ &\quad - \mu(\overline{y}_1, \overline{y}_0) + \iint E[BP_M] d\mu_c(y_1, y_0) \text{ and} \\ \theta_L &= E[\theta_L(X)] = E[\mu(Y_1, \overline{y}_0)] + E[\mu(\overline{y}_1, Y_0)] \qquad \text{(A.18)} \\ &\quad - \mu(\overline{y}_1, \overline{y}_0) + \iint E[BX_M] d\mu_c(y_1, y_0). \end{aligned}$$

Note that for given y_0 and y_1, $E[F_{jo}(y_j|X)|p(X)] = F_{jo}(y_j|p(X))$ $(j = 1, 0)$ and thus

$$E[BP_M] - E[BX_M] = F_P^L(y_1, y_0) - F^L(y_1, y_0).$$

By using the fact that $F_P^L(y_1, y_0) \leq F^L(y_1, y_0)$ for all (y_1, y_0), we have $E[BP_M] \leq E[BX_M]$, implying by comparing (A.17) and (A.18) that $\theta_{LP} \leq \theta_L$ and $\theta_{LP} = \theta_L$ iff $\iint [F_P^L(y_1, y_0) - F^L(y_1, y_0)] d\mu_c(y_1, y_0) = 0$. For a strict super-modular function $\mu(\cdot, \cdot)$, implying that any rectangle in (y_1, y_0)-plane has a positive μ_c measure, we obtain that $\theta_{LP} = \theta_L$ iff $F_P^L(y_1, y_0) = F^L(y_1, y_0)$ for μ_c-almost all (y_1, y_0). Moreover, from the proof of (A.14), we see that $F_P^L(y_1, y_0) = F^L(y_1, y_0)$ for μ_c-almost all (y_1, y_0) if and only if $\Pr\{F_{1o}(y_1|X) + F_{0o}(y_0|X) - 1 > 0|p(X)\} \in \{0, 1\}$ for μ_c-almost all (y_1, y_0). ■

Proof of Proposition 4.2: The proof is similar to that of Proposition 4.1.

Appendix B: Inference for Super-Modular μ in the Selection-on-Observables Framework

We have provided a comprehensive study of partial identification of θ_o under various scenarios in the paper. In this Appendix, we illustrate feasibility of inference by constructing confidence sets (CSs) for θ_o and its conditional version $\theta_o(x)$ for strict super-modular functions μ under the selection-on-observables assumption, i.e., Assumption (IX), which implies that Assumption (IC) holds with $X^* = X$. Technical proofs in this Appendix are relegated to the on-line Supplementary Appendices. Asymptotically valid inference procedures for θ_o and $\theta_o(x)$ corresponding to functions μ in other cases studied in the paper including latent threshold-crossing models remain to be developed.

Throughout this Appendix, we use $\Longrightarrow$ to denote weak convergence. All the limits are taken as the sample size goes to ∞.

B.1 Estimators of the Bounds and Assumptions

Suppose $\mu(\cdot,\cdot)$ is strict super-modular and right continuous. Let $Q_j(u|x) = F_{jo}^{-1}(u|x)$, $j = 0, 1$ and

$$\theta_L(x) = \int_0^1 \mu\left(Q_1(u|x), Q_0(1-u|x)\right) du,\ \theta_U(x) = \int_0^1 \mu\left(Q_1(u|x), Q_0(u|x)\right) du.$$

An application of CSS conditional on the covariate implies that $\theta_o(x)$ is partially identified: $\theta_L(x) \leq \theta_o(x) \leq \theta_U(x)$ and $\theta_L(x) = \theta_U(x)$ if and only if at least one of the conditional marginal distributions $F_{1o}(\cdot|x), F_{0o}(\cdot|x)$ is degenerate.

Suppose a random sample $\{Y_i, X_i, D_i\}_{i=1}^n$ on $\{Y, X, D\}$ is available. We estimate the conditional quantile function $Q_j(u|x)$ of Y given $X = x$ and $D = j$ using the local polynomial approach. Let

$$\ell_u(t) = t\left(u - I(t \leq 0)\right), \quad u \in [0,1]$$

be the quantile check function and $Y_{(1)} = \min_{i=1,\ldots,n} Y_i$, $Y_{(n)} = \max_{i=1,\ldots,n} Y_i$. Consider a kernel function $K(\cdot)$, a bandwidth $a_n > 0$, and an integer $s \geq 1$. Let $x = (x_1, \ldots, x_d)$ and $P_1(x)$ be the vector which stacks the power

$$x_1^{j_1} \times \cdots \times x_d^{j_d}, \quad 1 \leq j_1 + \cdots + j_d \leq s-1,$$

according to the lexicographic order. Define also $P(x) = \left(1, P_1(x)'\right)'$. The local polynomial estimator of $Q_j(u|x)$, j = 0,1, is defined as $\widehat{Q}_j(u|x) = \widehat{b}_{0j}(u|x)$, where $\widehat{b}_{0j}(u|x)$ and $\widehat{b}_{1j}(u|x)$ achieve the minimum of

$$\sum_{i=1}^n \ell_u\left(Y_i - b_0 - P_1(X_i - x)' b_1\right) I\{D_i = j\} \frac{1}{a_n^d} K\left(\frac{X_i - x}{a_n}\right), \quad b_0 \in \left[Y_{(1)}, Y_{(n)}\right],$$

where an appropriate convention is used to break ties.

The estimators of $\theta_L(x)$, $\theta_U(x)$, θ_L and θ_U are,

$$\widehat{\theta}_L(x) = \int_0^1 \mu\left(\widehat{Q}_1(u|x), \widehat{Q}_0(1-u|x)\right) du, \quad \widehat{\theta}_U(x) = \int_0^1 \mu\left(\widehat{Q}_1(u|x), \widehat{Q}_0(u|x)\right) du,$$

$$\widehat{\theta}_L = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_L\left(X_i\right), \quad \widehat{\theta}_U = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_U\left(X_i\right).$$

The restriction that $\widehat{Q}_j\left(u|x\right) = \widehat{b}_{0j}\left(u|x\right) \in \left[Y_{(1)}, Y_{(n)}\right]$ is useful in the extreme cases: $u = 0$ or $u = 1$. As discussed in Hall and van Keilegom (2009), for $u = 0{,}1$, the minimizers $\widehat{b}_{0j}\left(0|x\right)$ and $\widehat{b}_{0j}\left(1|x\right)$ may become infinite. The restriction that $\widehat{Q}_j\left(u|x\right) = \widehat{b}_{0j}\left(u|x\right) \in \left[Y_{(1)}, Y_{(n)}\right]$ is a sample version of a basic property of the population conditional quantile $Q_j\left(u|x\right)$ which lies between the minimal and maximal values taken by Y. Imposing these restrictions helps to get consistent $\widehat{\theta}_L\left(x\right)$ and $\widehat{\theta}_U\left(x\right)$.

We assume that the support of X given $D = j$ is the same as that of X denoted as $\mathcal{X}$. Let x be any point in $\mathcal{X}$ including its boundary. To establish the asymptotic distribution of $\left(\widehat{\theta}_L\left(x\right), \widehat{\theta}_U\left(x\right)\right)'$ and $\left(\widehat{\theta}_L, \widehat{\theta}_U\right)'$, we introduce the following assumptions. Let $p_j\left(x\right) = \Pr(D = j|x)$ and $f_j(y|x) \equiv \partial F_j(y|x)/\partial y$, where $F_j(y|x) \equiv F_{jo}(y|x)$ is the conditional cumulative distribution function of Y given $X = x$ and $D = j$, with support $\mathcal{S}_j = \{(x, y)\,; x \in \mathcal{X}, y \in \mathcal{Y}_j\left(x\right) \equiv \left[Q_j\left(0|x\right), Q_j\left(1|x\right)\right]\}$. Note that $\mathcal{Y}_j = [\inf_{x\in\mathcal{X}} Q_j\left(0|x\right), \sup_{x\in\mathcal{X}} Q_j\left(1|x\right)]$.

(A1) (i) The partial derivatives of $F_j\left(y|x\right)$ w.r.t. to x up to order s are continuous over $\mathcal{S}_j$, (ii) $\mathcal{S}_j$ is compact, $f_j(\cdot|\cdot)$ is continuously differentiable over $\mathcal{S}_j$ and satisfies: $\inf_{(y,x)\in\mathcal{S}_j} f_j\left(y|x\right) > 0$.

(A2) (i) $X|D = j$ is continuous with continuous probability density functions $f_j\left(\cdot\right)$ satisfying $\inf_{x\in\mathcal{X}} f_j\left(x\right) > 0$, $j = 0, 1$. Further, $p\left(\cdot\right) \in (0, 1)$ is continuous over $\mathcal{X}$, (ii) There is some $\kappa > 0$ such that, for all $\epsilon > 0$ small enough, any $x \in \mathcal{X}$, there is $x' \in \mathcal{X}$ such that

$$\mathcal{B}\left(x', \kappa\epsilon\right) \subset \mathcal{B}\left(x, \epsilon\right) \cap \mathcal{X},$$

where $\mathcal{B}\left(x, \epsilon\right)$ is the Euclidean ball with center x and radius ϵ.

(A3) $\mu\left(y_1, y_0\right)$ is twice differentiable on $\mathcal{Y}_1 \times \mathcal{Y}_0$ with bounded second-order partial derivatives.

(A4) (i) The kernel $K\left(\cdot\right)$ is non negative and Lipschitz, i.e., $|K\left(x\right) - K\left(x'\right)| \leq L\left\|x - x'\right\|$ for any x, $x' \in R^d$. The kernel $K\left(\cdot\right)$ has a compact support and is bounded away from 0 over the unit ball $\mathcal{B}\left(0, 1\right)$, (ii) The bandwidth sequence a_n satisfies $a_n \to 0$, $na_n^{d+s}/\log^3 n \to \infty$, and $na_n^{2s+d} \to 0$.

Assumption (A1)-(i) implies that the conditional quantile functions $Q_j\left(u|x\right)$, $j = 0, 1$, are continuously differentiable with respect to (x, u) up to order s. An important implication of Assumptions (A1)-(ii) and (A2) is that the quantile density function $1/\left(f_j\left(Q_j\left(u|x\right)|x\right) f_j\left(x\right)\right)$, which is proportional to the asymptotic variance of many nonparametric quantile estimators, stays bounded away from infinity. Assumption (A2)-(ii), which is from Fan and Guerre (2016), ensures that the bias of the local polynomial quantile estimators $\widehat{Q}_j\left(u|x\right)$ is of order $O\left(a_n^s\right)$ including for x on the boundary of $\mathcal{X}$ and also u close to 0 and 1, see Proposition C.3 in online Appendix C. The other assumptions are standard, except the condition $na_n^{d+s}/\log^3 n \to \infty$ in Assumption (A4)-(ii). This condition is used to establish the asymptotic normality of $\widehat{\theta}_L$ and $\widehat{\theta}_U$ and is briefly discussed after Theorem B.2.

B.2 Asymptotic Normality

Let $e_0 = (1, 0, \ldots, 0)'$ denote the first vector of the canonical basis and

$$V^2_{K,a_n}(x) = e_0' \left[\int P(v) P(v)' K(v) 1(x + a_n v \in \mathcal{X}) dv \right]^{-1} \left[\int P(v) P(v)' K^2(v) 1(x + a_n v \in \mathcal{X}) dv \right] \times \left[\int P(v) P(v)' K(v) 1(x + a_n v \in \mathcal{X}) dv \right]^{-1} e_0.$$

Lemma C.2 in online Appendix C shows that under Assumption (A2)-(ii), $V^2_{K,a_n}(x)$ is well-defined uniformly over the support $\mathcal{X}$ provided that a_n is small enough. Define also, for $\mu_j(y_1, y_0) = \partial \mu(y_1, y_0) / \partial y_j$,

$$G_{0L}(u) = \frac{\mu_0(Q_1(u|x), Q_0(1-u|x))}{f_0(Q_0(1-u|x)|x)}, \quad G_{0U}(u) = \frac{\mu_0(Q_1(u|x), Q_0(u|x))}{f_0(Q_0(u|x)|x)},$$
$$G_{1L}(u) = \frac{\mu_1(Q_1(u|x), Q_0(1-u|x))}{f_1(Q_1(u|x)|x)}, \quad G_{1U}(u) = \frac{\mu_1(Q_1(u|x), Q_0(u|x))}{f_1(Q_1(u|x)|x)}.$$

We are now ready to state the joint asymptotic normality of $\left(\widehat{\theta}_L(x), \widehat{\theta}_U(x)\right)'$.

THEOREM B.1 *Suppose Assumption (IX) and (A1)-(A4) hold. Then, for any $x \in \mathcal{X}$,*

$$\frac{\sqrt{na_n^d}}{V_{K,a_n}(x)} \begin{pmatrix} \widehat{\theta}_L(x) - \theta_L(x) \\ \widehat{\theta}_U(x) - \theta_U(x) \end{pmatrix} \Longrightarrow N\left[0, \begin{pmatrix} \sigma^2_L(x) & \sigma_{LU}(x) \\ \sigma_{LU}(x) & \sigma^2_U(x) \end{pmatrix}\right],$$

with

$$\sigma^2_L(x) = \int_0^1 \int_0^1 \frac{G_{0L}(u) G_{0L}(v)}{f_0(x) \Pr(D=0)} \{\min(1-u, 1-v) - (1-u)(1-v)\} du dv + \int_0^1 \int_0^1 \frac{G_{1L}(u) G_{1L}(v)}{f_1(x) \Pr(D=1)} \{\min(u, v) - uv\} du dv,$$
$$\sigma^2_U(x) = \int_0^1 \int_0^1 \left\{ \frac{G_{0U}(u) G_{0U}(v)}{f_0(x) \Pr(D=0)} + \frac{G_{1U}(u) G_{1U}(v)}{f_1(x) \Pr(D=1)} \right\} \{\min(u, v) - uv\} du dv,$$

and

$$\sigma_{LU}(x) = \int_0^1 \int_0^1 \frac{G_{0L}(u) G_{0U}(v)}{f_0(x) \Pr(D=0)} \{\min(1-u, v) - (1-u) v\} du dv + \int_0^1 \int_0^1 \frac{G_{1L}(u) G_{1U}(v)}{f_1(x) \Pr(D=1)} \{\min(u, v) - uv\} du dv.$$

Theorem B.1 holds for all x in $\mathcal{X}$, including the boundaries of the support of the covariate X, showing in particular that $\left(\widehat{\theta}_L(x), \widehat{\theta}_U(x)\right)'$ is consistent when x lies on the boundary. The asymptotic normality stated in Theorem B.1 holds under the additional condition that the variance dominates the bias, that is $a_n^s = o\left(1/\sqrt{na_n^d}\right)$ as ensured by Assumption (A4)-(ii). The asymptotic variance of Theorem B.1 involves the partial derivatives of $\mu(y_1, y_0)$ due to the use of the Functional Delta method, the inverse of $f_j(Q_j(u|x)|x)$ which is typical of quantile estimation asymptotics, and the inverse of $f_j(x)$ as expected from a local polynomial method.

The proof of Theorem B.1 uses a Bahadur representation of the local linear quantile estimator $\widehat{Q}_j(u|x)$ which is also useful in other econometrics contexts, see Guerre and Sabbah (2012) and Kong, Linton and Xia (2010) among others. It is used here to derive sum approximations for $\widehat{\theta}_L$ and $\widehat{\theta}_U$,

$$\widehat{\theta}_L = \theta_L + \frac{1}{n}\sum_{i=1}^{n}\left(\theta_L(X_i) - E[\theta_L(X)] + r_L(W_i)\right) + O_P\left(a_n^s + \left(\frac{\log n}{na_n^d}\right)^{3/4}\right) + o_P\left(\frac{1}{\sqrt{n}}\right),$$
$$\widehat{\theta}_U = \theta_U + \frac{1}{n}\sum_{i=1}^{n}\left(\theta_U(X_i) - E[\theta_U(X)] + r_U(W_i)\right) + O_P\left(a_n^s + \left(\frac{\log n}{na_n^d}\right)^{3/4}\right) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

where

$$\begin{aligned}
r_L(W) &= \int_0^1 \frac{\mu_0\left(Q_1(u|X), Q_0(1-u|X)\right)}{\Pr(D=0) f_0\left(Q_0(1-u|X)|X\right)} 1(D=0)\left\{1\left(Y \le Q_0(1-u|X)\right) - (1-u)\right\} du \\
&\quad + \int_0^1 \frac{\mu_1\left(Q_1(u|X), Q_0(1-u|X)\right)}{\Pr(D=1) f_1\left(Q_1(u|X)|X\right)} 1(D=1)\left\{1\left(Y \le Q_1(u|X)\right) - u\right\} du, \\
r_U(W) &= \int_0^1 \frac{\mu_0\left(Q_1(u|X), Q_0(u|X)\right)}{\Pr(D=0) f_0\left(Q_0(u|X)|X\right)} 1(D=0)\left\{1\left(Y \le Q_0(u|X)\right) - u\right\} du \\
&\quad + \int_0^1 \frac{\mu_1\left(Q_1(u|X), Q_0(1-u|X)\right)}{\Pr(D=1) f_1\left(Q_1(u|X)|X\right)} 1(D=1)\left\{1\left(Y \le Q_1(u|X)\right) - u\right\} du.
\end{aligned}$$

This gives the next Theorem which states the asymptotic normality of $\left(\widehat{\theta}_L, \widehat{\theta}_U\right)$.

THEOREM B.2 *Suppose Assumption (IX) and (A1)-(A4) hold with* $na_n^{2s} = o(1)$ *and* $na_n^{3d}/\log^3 n \to \infty$. *Then*

$$\sqrt{n}\begin{pmatrix} \widehat{\theta}_L - \theta_L \\ \widehat{\theta}_U - \theta_U \end{pmatrix} \Longrightarrow N(0, \Sigma), \text{ where } \Sigma = \mathrm{Var}\begin{pmatrix} \theta_L(X_i) + r_L(W_i) \\ \theta_U(X_i) + r_U(W_i) \end{pmatrix}.$$

Compared to Theorem B.1, Theorem B.2 includes two additional bandwidth conditions, $na_n^{2s} = o(1)$ and $na_n^{3d}/\log^3 n \to \infty$, which ensure that the remainder terms in the bias and Bahadur expansions and of the local polynomial estimators $\widehat{Q}_j(u|x)$ are negligible with respect to $1/\sqrt{n}$. Note that these two conditions implicitly impose the smoothness condition $s > 3d/2$, suggesting that the order of the local polynomial quantile estimators and the order of differentiability of the conditional quantile function must increase with the dimension of the covariate X. This is in line with the qualitatively similar dependence condition needed in Powell, Stock and Stoker (1989, Theorem 3.3) for average derivative estimation. Whether the coefficient 3/2 in front of the dimension is outside the scope of the present paper.

B.3 Variance Estimation and Asymptotic Inference

The asymptotic variance of $\left(\widehat{\theta}_L(x), \widehat{\theta}_U(x)\right)'$ can be estimated by plugging in the expression of $\sigma_L^2(x)$, $\sigma_U^2(x)$, $\sigma_{LU}(x)$ estimators of $\Pr(D=j)$, $f_j(x)$, and $q_j(u|x) = 1/f_j\left(Q_j(u|x)|x\right)$, $j = 0, 1$ as introduced now. Consider a univariate symmetric and Lipchitz kernel $K_1(y)$ with a compact support and $\int K_1(y)\,dy = 1$

and a bandwidth $a_{1n} < 1/2$. Define

$$\widehat{q}_j(u|x) = \begin{cases} \frac{\widehat{Q}_j(u+a_{1n}|x)-\widehat{Q}_j(u|x)}{a_{1n}}, & u \in [0,1/2], \\ \frac{\widehat{Q}_j(u|x)-\widehat{Q}_j(u-a_{1n}|x)}{a_{1n}}, & u \in (1/2,1], \end{cases}$$

$$\widehat{f}_j(x) = \frac{\sum_{i=1}^n 1(D_i = j) K\left(\frac{X_i - x}{a_n}\right)}{a_n^d \sum_{i=1}^n 1(D_i = j)}, \text{ and}$$

$$\widehat{\Pr(D=j)} = \frac{1}{n}\sum_{i=1}^n 1(D_i = j), \quad j = 0, 1.$$

is a modification of an estimator of $\partial Q_j(u|x)/\partial u$ in Guerre and Sabbah (2012), see also Hall and Sheather (1988) for an unconditional version, which is well defined near the boundaries $u = 0, 1$. The idea behind $\widehat{q}_j(u|x)$ is that Newton's difference quotient is an estimator of the derivative

$$q_j(u|x) = \frac{\partial Q_j(u|x)}{\partial u} = \frac{1}{f_j(Q_j(u|x)|x)}.$$

As $\widehat{Q}_j(u|x)$ is consistent for all x in $\mathcal{X}$, $\widehat{q}_j(u|x)$ is a consistent estimator of $q_j(u|x)$ even when x lies in the boundaries of $\mathcal{X}$ and u is close to the boundary $u = 0, 1$. This will hold provided a_{1n} is negligible with respect to the consistency rate of $\widehat{Q}_j(u|x)$ as assumed in the results below.

Let $\widehat{\sigma}_L^2(x)$ and $\widehat{\sigma}_U^2(x)$ be the corresponding plug-in estimators of $\sigma_L^2(x)$ and $\sigma_U^2(x)$. It follows from Theorem 3.1 that $\theta_L(x) = \theta_U(x)$ if and only if at least one of $F_1(\cdot|x), F_0(\cdot|x)$ is degenerate. As Assumption (A1) excludes the case that at least one of $F_1(\cdot|x), F_0(\cdot|x)$ is degenerate, we only need to consider the case $\theta_U(x) > \theta_L(x)$. Following Horowitz and Manski (2000), define the confidence set

$$\widehat{CS}_{1-\alpha}(x) = \left[\widehat{\theta}_L(x) - \frac{\widehat{\sigma}_L(x)}{\sqrt{na_n^d}} z_{1-\alpha}, \widehat{\theta}_U(x) + \frac{\widehat{\sigma}_U(x)}{\sqrt{na_n^d}} z_{1-\alpha}\right],$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution. The next Theorem shows that $\widehat{CS}_{1-\alpha}(x)$ contains the true $\theta_o(x)$ with an asymptotic probability of $1-\alpha$.

THEOREM B.3 *Suppose the conditions of Theorem B.1 hold with* $\left(\log n/\left(na_n^d\right)\right)^{1/2} = o(a_{1n})$ *and* $0 < \alpha < 1$. *Then for any* x *in the interior of* $\mathcal{X}$,

$$\lim_{n\to\infty} \inf_{\theta_L(x) \le \theta_0(x) \le \theta_U(x)} \Pr\left(\theta_o(x) \in \widehat{CS}_{1-\alpha}(x)\right) = 1 - \alpha.$$

Compared to Theorem B.1, Theorem B.3 does not allow for x to lie on the boundary of the support $\mathcal{X}$. This is because $\widehat{f}_j(x)$ is a biased estimator of $f_j(x)$ for those x. Finding a bias correction for $\widehat{f}_j(x)$ is in principle feasible using an estimation of the support $\mathcal{X}$. By contrast it is possible to find a confidence interval for θ_o which is not affected by such issues as detailed now.

Estimation of the asymptotic variance of $\left(\widehat{\theta}_L, \widehat{\theta}_U\right)'$ can be done by plugging $\widehat{Q}_j(u|x)$, $\widehat{q}_j(u|x)$ and $\widehat{\Pr(D=j)}$ in the expression of $r_L(w)$ and $r_U(w)$ to obtain some estimators $\widehat{r}_L(w)$ and $\widehat{r}_U(w)$ of these functions. A natural estimator of Σ is then

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \widehat{\theta}_L(X_i) + \widehat{r}_L(W_i) - \left(\widehat{\theta}_L + \overline{\widehat{r}_L(W)}\right) \\ \widehat{\theta}_U(X_i) + \widehat{r}_U(W_i) - \left(\widehat{\theta}_U + \overline{\widehat{r}_U(W)}\right) \end{pmatrix} \begin{pmatrix} \widehat{\theta}_L(X_i) + \widehat{r}_L(W_i) - \left(\widehat{\theta}_L + \overline{\widehat{r}_L(W)}\right) \\ \widehat{\theta}_U(X_i) + \widehat{r}_U(W_i) - \left(\widehat{\theta}_U + \overline{\widehat{r}_U(W)}\right) \end{pmatrix}'.$$

Let $\widehat{\sigma}_L^2$, $\widehat{\sigma}_U^2$, and $\widehat{\sigma}_{LU}$ be the entries of $\widehat{\Sigma}$. Then a confidence set for θ_o is

$$\widehat{CS}_{1-\alpha} = \left[\widehat{\theta}_L - \frac{\widehat{\sigma}_L}{\sqrt{n}} z_{1-\alpha}, \widehat{\theta}_U + \frac{\widehat{\sigma}_U}{\sqrt{n}} z_{1-\alpha}\right].$$

THEOREM B.4 *Suppose the conditions of Theorem B.2 hold with* $na_{1n}^{d+1}/\log n \to +\infty$, $\log n/\left(na_n^d\right) = o\left(a_{1n}^s\right)$, *and* $0 < \alpha < 1$. *Then*

$$\lim_{n\to\infty} \inf_{\theta_L \leq \theta_0 \leq \theta_U} \Pr\left(\theta_o \in \widehat{CS}_{1-\alpha}\right) = 1 - \alpha.$$

Both Theorem B.3 and Theorem B.4 are pointwise results in the true probability measure characterizing the population. To construct asymptotically uniformly valid CSs, we could follow Imbens and Manski (2004) and Stoye (2009). To do so, we need to allow for at least one of $F_1\left(\cdot|x\right), F_0\left(\cdot|x\right)$ to be degenerate and strengthen the asymptotic distribution results so that they hold uniformly over a class of distributions generating the sample information. This could be done at the cost of increased technical complexity.

Appendix C: Algebraic Derivations for Examples (i)-(IC) and (i)-(IU)

This Appendix is self-contained. It presents Examples (i)-(IC) and (i)-(IU) with detailed algebraic derivations.

Example (i)-(IC) (Correlation Coefficient). Let the covariate X^* be univariate. For notational simplicity, we denote X^* as X in this example. Suppose the distribution of (Y_j, X) is known to be a bivariate normal distribution:

$$\begin{pmatrix} Y_j \\ X \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_j^2 & \sigma_j\rho_{jX} \\ \sigma_j\rho_{jX} & 1 \end{pmatrix}\right],\ j = 0, 1.$$

Then Assumption (IC) is satisfied with $Y_j|X = x \sim N\left(\sigma_j\rho_{jX}x, \sigma_j^2\left(1-\rho_{jX}^2\right)\right)$, $j = 0, 1$, and $X \sim N(0, 1)$. Obviously, $Y_j \sim N\left(0, \sigma_j^2\right)$. Suppose $\sigma_j^2 > 0$, $j = 1, 0$. Using Theorem 2 in Cambanis, Simons, and Stout (1976), we get $\rho_{10} \equiv corr(Y_1, Y_0) \in \left[\rho^L, \rho^U\right] = [-1, 1]$, so ρ_{10} is not identified. Now, we know that $Y_j|X = x \sim N\left(\sigma_j\rho_{jX}x, \sigma_j^2\left(1-\rho_{jX}^2\right)\right)$ and $X \sim N(0, 1)$. Theorem 3.2 (i) yields: $\rho_L \leq \rho_{10} \leq \rho_U$, where

$$\rho_L = \rho_{0X}\rho_{1X} - \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)} \text{ and } \rho_U = \rho_{0X}\rho_{1X} + \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}.$$

Three conclusions are immediate. First, $\rho^L \leq \rho_L$, $\rho_U \leq \rho^U$, and at least one of the inequalities holds as a *strict* inequality if and only if $\rho_{0X} + \rho_{1X} \neq 0$ or $\rho_{0X} - \rho_{1X} \neq 0$, implying that $[\rho_L, \rho_U] = [-1, 1]$ iff X is independent of (Y_1, Y_0). This conclusion is consistent with Theorem 3.3, since we can show that

$$\Pr(F_{1o}(y_1|X) + F_{0o}(y_0|X) - 1 > 0) \in \{0, 1\} \text{ for all } (y_1, y_0) \text{ iff } \rho_{0X} + \rho_{1X} = 0 \text{ and}$$
$$\Pr(F_{1o}(y_1|X) - F_{0o}(y_0|X) < 0) \in \{0, 1\} \text{ for all } (y_1, y_0) \text{ iff } \rho_{0X} - \rho_{1X} = 0.$$

In fact, noting that $F_{jo}(y_j|X) = \Phi[(y_j - \sigma_j\rho_{jX}X)/\sigma_j(1-\rho_{jX}^2)^{1/2}]$ $(j = 1, 0)$, where Φ is the cdf of $N(0, 1)$, we conclude that for the lower bound, $F_{1o}(y_1|X) + F_{0o}(y_0|X) - 1 > 0$ is equivalent to

$$\Phi\left[\frac{y_1 - \sigma_1\rho_{1X}X}{\sigma_1\sqrt{1-\rho_{1X}^2}}\right] > \Phi\left[-\frac{y_0 - \sigma_0\rho_{0X}X}{\sigma_0\sqrt{1-\rho_{0X}^2}}\right] \Leftrightarrow \sum_{j=0,1}\frac{y_j}{\sigma_j\sqrt{1-\rho_{jX}^2}} > \left[\sum_{j=0,1}\frac{\rho_{jX}}{\sqrt{1-\rho_{jX}^2}}\right]X.$$

It follows from $X \sim N(0, 1)$ that $\Pr(F_{1o}(y_1|X) + F_{0o}(y_0|X) - 1 > 0) \in \{0, 1\}$ for all (y_1, y_0) if and only if $\rho_{1X}/\sqrt{1-\rho_{1X}^2} + \rho_{0X}/\sqrt{1-\rho_{0X}^2} = 0$, which is a condition equivalent to $\rho_{0X} + \rho_{1X} = 0$. Similarly, we can show the result for the upper bound. Second, when $\rho_{0X}\rho_{1X} > 0$ and $\rho_{0X}^2 + \rho_{1X}^2 > 1$, we have $0 < \rho_L \leq \rho_U$, so ρ_{10} is positive and when $\rho_{0X}\rho_{1X} < 0$ and $\rho_{0X}^2 + \rho_{1X}^2 > 1$, we have $\rho_L \leq \rho_U < 0$, so ρ_{10} is negative. Third, ρ_{10} is point identified (i.e., $\rho_L = \rho_U = \rho_{10}$) if and only if $\rho_{0X}^2 = 1$ or $\rho_{1X}^2 = 1$; this condition is equivalent to $Var[Y_0|X = x] = 0$ or $Var[Y_1|X = x] = 0$ for all x, that is, at least one of the conditional marginal distributions of Y_0 and Y_1 given $X = x$ is degenerate (at $\sigma_0\rho_{0X}x$ and $\sigma_1\rho_{1X}x$ respectively) for almost all x; in this case, ρ_{10} is point identified at either $\rho_{0X}sign(\rho_{1X})$ or $\rho_{1X}sign(\rho_{0X})$. The third conclusion confirms that in Theorem 3.2 (ii).

Example (i)-(IC) demonstrates that when the dependence between (Y_1, Y_0) and covariate X is strong enough in the sense that $\rho_{0X}^2 + \rho_{1X}^2 > 1$, the identified set for ρ_{10} excludes 0 so identifies the sign of ρ_{10}.

Example (i)-(IU) (Correlation Coefficient). Consider the following special case of the latent threshold-crossing model (3):

$$Y_1 = g_1(X) + U_1,\ Y_0 = g_0(X) + U_0,\ \text{and}\ D = I\{g(Z) - \epsilon > 0\}.$$

Since the distribution of ϵ conditional on X is normalized to be $U(0,1)$, the distribution of $V \equiv \Phi^{-1}(\epsilon)$ conditional on X is $N(0,1)$, where $\Phi(\cdot)$ is the cdf of $N(0,1)$. Suppose that $(U_1, U_0, \epsilon)'$ is independent of Z conditional on X, implying that Assumptions (IU)-(ii) holds. Then the joint distribution of $(U_1,U_0,V,X,Z)'$ can be expressed as $f(u_1,u_0,v,x,z) = f(u_1,u_0,v,x)f(z|x)$. Thus we only need to consider the joint distribution of $(U_1,U_0,V,X)'$.

Let $U = (U_1,U_0)'$, $X^* = (V,X)'$ and assume for simplicity that $g_i(X) = \mu_i$ $(i = 1,0)$ are constants and $(U_1,U_0,V,X)'$ follows a multivariate normal distribution:

$$\begin{pmatrix} U \\ X^* \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right], \tag{C.1}$$

where $\Sigma_{21} = \Sigma_{12}'$,

$$\Sigma_{11} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_0\rho_{10} \\ \sigma_1\sigma_0\rho_{10} & \sigma_0^2 \end{pmatrix},\ \Sigma_{12} = \begin{pmatrix} \sigma_1\rho_{1V} & \sigma_1\sigma_X\rho_{1X} \\ \sigma_0\rho_{0V} & \sigma_0\sigma_X\rho_{0X} \end{pmatrix},\ \text{and}\ \Sigma_{22} = \begin{pmatrix} 1 & \sigma_X\rho_{XV} \\ \sigma_X\rho_{XV} & \sigma_X^2 \end{pmatrix}.$$

Then the conditional distribution of $Y \equiv (Y_1,Y_0)'$ given X^* is normal:

$$Y|X^* \sim N\left(\mu + \Sigma_{12}\Sigma_{22}^{-1}X^*, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right), \tag{C.2}$$

where $\mu = (\mu_1,\mu_0)'$ and the expression for $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ is given as follows:

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} a_{11} & a_{10} \\ a_{10} & a_{00} \end{pmatrix}, \tag{C.3}$$

in which

$$\begin{aligned}
a_{11} &= \sigma_1^2\left(\frac{1-\rho_{XV}^2-\rho_{1V}^2-\rho_{1X}^2+2\rho_{1V}\rho_{1X}\rho_{XV}}{1-\rho_{XV}^2}\right), \\
a_{00} &= \sigma_0^2\left(\frac{1-\rho_{XV}^2-\rho_{0V}^2-\rho_{0X}^2+2\rho_{0V}\rho_{0X}\rho_{XV}}{1-\rho_{XV}^2}\right),\ \text{and} \\
a_{10} &= \sigma_1\sigma_0\left(\rho_{10} - \frac{\rho_{1V}\rho_{0V}+\rho_{0X}\rho_{1X}-\rho_{0V}\rho_{1X}\rho_{XV}-\rho_{0X}\rho_{1V}\rho_{XV}}{1-\rho_{XV}^2}\right).
\end{aligned}$$

The fact that $-1 \le corr(Y_1,Y_0|X^*) = a_{10}/\sqrt{a_{11}a_{00}} \le 1$ implies that $\rho_L \le corr(Y_1,Y_0) \equiv \rho_{10} \le \rho_U$, where

$$\begin{aligned}
\rho_L = &\left(\frac{\rho_{1V}\rho_{0V}+\rho_{0X}\rho_{1X}-\rho_{0V}\rho_{1X}\rho_{XV}-\rho_{0X}\rho_{1V}\rho_{XV}}{1-\rho_{XV}^2}\right) - \\
&\frac{\sqrt{(1-\rho_{XV}^2-\rho_{1V}^2-\rho_{1X}^2+2\rho_{1V}\rho_{1X}\rho_{XV})(1-\rho_{XV}^2-\rho_{0V}^2-\rho_{0X}^2+2\rho_{0V}\rho_{0X}\rho_{XV})}}{1-\rho_{XV}^2}\ \text{and}
\end{aligned} \tag{C.4}$$

$$\begin{aligned}
\rho_U = &\left(\frac{\rho_{1V}\rho_{0V}+\rho_{0X}\rho_{1X}-\rho_{0V}\rho_{1X}\rho_{XV}-\rho_{0X}\rho_{1V}\rho_{XV}}{1-\rho_{XV}^2}\right) + \\
&\frac{\sqrt{(1-\rho_{XV}^2-\rho_{1V}^2-\rho_{1X}^2+2\rho_{1V}\rho_{1X}\rho_{XV})(1-\rho_{XV}^2-\rho_{0V}^2-\rho_{0X}^2+2\rho_{0V}\rho_{0X}\rho_{XV})}}{1-\rho_{XV}^2}.
\end{aligned} \tag{C.5}$$

Case I. Suppose U_1 and U_0 are jointly independent of V conditional on X, Z. Then the selection-on-observables assumption (i.e., Assumption (IX)) holds. It follows from Assumptions (IU)-(ii) that U_1 and U_0 are also jointly independent of V conditional on X, implying that $\rho_{1V} - \rho_{1X}\rho_{XV} = 0$ and $\rho_{0V} - \rho_{0X}\rho_{XV} = 0$. Both constraints follow from the fact that $U^* \equiv (U_1, U_0, V)'|X \sim N(\Sigma_{U^*X}X/\sigma_X^2, \Sigma_{U^*} - \Sigma_{U^*X}\Sigma_{XU^*}/\sigma_X^2)$, where $\Sigma_{XU^*} = \Sigma'_{U^*X} = \begin{pmatrix} \sigma_1\sigma_X\rho_{1X} & \sigma_0\sigma_X\rho_{0X} & \sigma_X\rho_{XV} \end{pmatrix}$,

$$\Sigma_{U^*} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_0\rho_{10} & \sigma_1\rho_{1V} \\ & \sigma_0^2 & \sigma_0\rho_{0V} \\ & & 1 \end{pmatrix}, \text{ and}$$

$$\Sigma_{U^*} - \Sigma_{U^*X}\Sigma_{XU^*}/\sigma_X^2 = \begin{pmatrix} \sigma_1^2\left(1-\rho_{1X}^2\right) & \sigma_1\sigma_0\left(\rho_{10}-\rho_{1X}\rho_{0X}\right) & \sigma_1\left(\rho_{1V}-\rho_{1X}\rho_{XV}\right) \\ & \sigma_0^2\left(1-\rho_{0X}^2\right) & \sigma_0\left(\rho_{0V}-\rho_{0X}\rho_{XV}\right) \\ & & 1-\rho_{XV}^2 \end{pmatrix}. \tag{C.6}$$

It follows from $\rho_{1V} - \rho_{1X}\rho_{XV} = 0$ and $\rho_{0V} - \rho_{0X}\rho_{XV} = 0$ that the bounds in (C.4) and (C.5) reduce to those in Example (i)-(IC):

$$\rho_L = \rho_L^{(1)} \equiv \rho_{0X}\rho_{1X} - \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)} \text{ and} \tag{C.7}$$

$$\rho_U = \rho_U^{(1)} \equiv \rho_{0X}\rho_{1X} + \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}. \tag{C.8}$$

It should be noted that the bounds in (C.7) and (C.8) are also those obtained by using only the conditional distribution information given X (that is, from $-1 \leq corr(Y_1, Y_0|X) \leq 1$, we can get $\rho_L^{(1)} \leq corr(Y_1, Y_0) \equiv \rho_{10} \leq \rho_U^{(1)}$).

Case II. We now demonstrate that when there is endogenous selection, i.e., U_1, U_0 are not jointly independent of V conditional on X, Z, the bounds in (C.7) and (C.8) may be tightened. Consider the special case of $\rho_{XV} = 0$. In this case, the bounds ρ_L and ρ_U in (C.4) and (C.5) reduce to:

$$\rho_L^{(2)} \equiv \left(\rho_{1V}\rho_{0V} + \rho_{0X}\rho_{1X}\right) - \sqrt{\left(1-\rho_{1V}^2-\rho_{1X}^2\right)\left(1-\rho_{0V}^2-\rho_{0X}^2\right)} \text{ and} \tag{C.9}$$

$$\rho_U^{(2)} \equiv \left(\rho_{1V}\rho_{0V} + \rho_{0X}\rho_{1X}\right) + \sqrt{\left(1-\rho_{1V}^2-\rho_{1X}^2\right)\left(1-\rho_{0V}^2-\rho_{0X}^2\right)}. \tag{C.10}$$

A straightforward calculation shows that (i) $\rho_L^{(1)} \leq \rho_L^{(2)}$ and $\rho_U^{(2)} \leq \rho_U^{(1)}$, implying that on the one hand, with endogenous selection (i.e., $\rho_{1V} \neq 0$ and $\rho_{0V} \neq 0$) the identified set would be tightened; on the other hand, the identified set based on more conditional distribution information (i.e., given $X^* = (V, X)'$) should be smaller than that based on less conditional distribution information (i.e., given X only), (ii) $\rho_L^{(1)} = \rho_L^{(2)}$ iff $\rho_{1V}\sqrt{1-\rho_{0X}^2} + \rho_{0V}\sqrt{1-\rho_{1X}^2} = 0$ and $\rho_U^{(2)} = \rho_U^{(1)}$ iff $\rho_{1V}\sqrt{1-\rho_{0X}^2} - \rho_{0V}\sqrt{1-\rho_{1X}^2} = 0$.

The result (ii) can also be obtained from Proposition 4.2. Here we show it only for the case of $\rho_U^{(2)} = \rho_U^{(1)}$. Note from (C.2) that $Y_j|X^* \sim N\left[\mu_j^*, a_{jj}\right]$ and $F_{jo}\left(y_j|X^*\right) = \Phi\left[\left(y_j - \mu_j^*\right)/\sqrt{a_{jj}}\right]$ $(j = 1, 0)$, where $\mu_j^* \equiv \mu_j + [b_{jV}V + b_{jX}X]$ with $b_{jV} \equiv \sigma_j(\rho_{jV} - \rho_{jX}\rho_{XV})/(1-\rho_{XV}^2)$ and $b_{jX} \equiv \sigma_j(\rho_{jX} - \rho_{jV}\rho_{XV})/\sigma_X\left(1-\rho_{XV}^2\right)$, and a_{jj} $(j = 1, 0)$ are defined in (C.3). Then $F_{1o}\left(y_1|X^*\right) - F_{0o}\left(y_0|X^*\right) < 0$ is equivalent to $(y_1 - \mu_1^*)/\sqrt{a_{11}} < (y_0 - \mu_0^*)/\sqrt{a_{00}}$, specifically,

$$\frac{y_1-\mu_1}{\sqrt{a_{11}}} - \frac{y_0-\mu_0}{\sqrt{a_{00}}} < \left(\frac{b_{1V}}{\sqrt{a_{11}}} - \frac{b_{0V}}{\sqrt{a_{00}}}\right)V + \left(\frac{b_{1X}}{\sqrt{a_{11}}} - \frac{b_{0X}}{\sqrt{a_{00}}}\right)X. \tag{C.11}$$

Obviously, since $V|X \sim N(0,1)$ when $\rho_{XV} = 0$, $\Pr\left(F_{1o}\left(y_1|X^*\right) - F_{0o}\left(y_0|X^*\right) < 0|X\right) \in \{0,1\}$ for all (y_1, y_0) if and only if the coefficient of V in (C.11) is equal to zero, that is, $b_{1V}/\sqrt{a_{11}} - b_{0V}/\sqrt{a_{00}} = 0$. This condition can be reduced to $\rho_{1V}\sqrt{1-\rho_{0X}^2} - \rho_{0V}\sqrt{1-\rho_{1X}^2} = 0$ when $\rho_{XV} = 0$. In conclusion, with $\rho_{XV} = 0$ and $\rho_{jX}^2 < 1$, as long as U_1 and U_0 are not jointly independent of V conditional on X (i.e., $\rho_{1V} \neq 0$ or $\rho_{0V} \neq 0$), the identified set can be strictly tightened.

The bounds $\rho_L^{(2)}$ and $\rho_U^{(2)}$ with endogenous selection are able to identify the sign of ρ_{10} under quite weak conditions on the dependence between (Y_1, Y_0) and the observable covariate X. It follows from (C.9) and (C.10) that when $\rho_{1V}\rho_{0V} + \rho_{0X}\rho_{1X} > 0$ and

$$\rho_{1X}^2\left(1-\rho_{0V}^2\right) + \rho_{0X}^2\left(1-\rho_{1V}^2\right) + 2\rho_{1V}\rho_{0V}\rho_{0X}\rho_{1X} + \rho_{0V}^2 + \rho_{1V}^2 - 1 > 0, \tag{C.12}$$

we have $0 < \rho_L^{(2)} \leq \rho_U^{(2)}$, implying that ρ_{10} is positive; when $\rho_{1V}\rho_{0V} + \rho_{0X}\rho_{1X} < 0$ and (C.12) holds, we have $\rho_L^{(2)} \leq \rho_U^{(2)} < 0$, implying a negative ρ_{10}. From (C.12), we can see that as long as the correlations between $V \equiv \Phi^{-1}(\epsilon)$ and U_j (i.e., ρ_{1V} and ρ_{0V}) are strong enough so that $\rho_{0V}^2 + \rho_{1V}^2 > 1$, we can identify the sign of ρ_{10} under quite weak conditions on ρ_{0X} and ρ_{1X}: $\rho_{10} > 0$ when $\rho_{0X}\rho_{1X} \geq 0$ with $\rho_{1V}\rho_{0V} > 0$, and $\rho_{10} < 0$ when $\rho_{0X}\rho_{1X} \leq 0$ with $\rho_{1V}\rho_{0V} < 0$. Obviously, these conditions on ρ_{0X} and ρ_{1X} (i.e., $\rho_{0X}\rho_{1X} \geq 0$ or $\rho_{0X}\rho_{1X} \leq 0$) cannot identify the sign of ρ_{10} without endogenous selection.

References

[1] Abbring, J. H., Heckman, J., 2007. Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation, in *Handbook of Econometrics* 6B, 5145-5301.

[2] Cambanis, S., Simons, G., Stout, W., 1976. Inequalities for $\mathcal{E}k(X,Y)$ when the Marginals are Fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 36, 285-294.

[3] Carneiro, P., Lee, S., 2009. Estimating Distributions of Potential Outcomes Using Instrumental Variables with an Application to Changes in College Enrolment and Wage Inequality. *Journal of Econometrics* 149, 191-208.

[4] Dehejia, R., Wahba, S., 1999. Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs. *Journal of the American Statistical Association* 94, 1053-1062.

[5] Embrechts, P., Hoeing, A., Juri, A., 2003. Using Copulae to Bound the Value-at-Risk for Functions of Dependent Risks. *Finance & Stochastics* 7(2), 145-167.

[6] Embrechts, P., Hoeing, A., Puccetti, G., 2005. Worst VaR Scenarios. *Insurance: Mathematics and Economics* 37, 115-134.

[7] Fan, Y., Guerre, E., 2016. Multivariate Local Polynomial Estimators: Uniform Boundary Properties and Asymptotic Linear Representation, forthcoming in *Advances in Econometrics*.

[8] Fan, Y., Park, S., 2009. Partial Identification of the Distribution of Treatment Effects and its Confidence Sets, in Thomas B. Fomby and R. Carter Hill (ed.) *Nonparametric Econometric Methods* (Advances in Econometrics, Volume 25), Emerald Group Publishing Limited, pp.3-70.

[9] Fan, Y., Park, S., 2010. Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference. *Econometric Theory* 26, 931-951.

[10] Fan, Y., Park, S., 2012. Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments. *Journal of Econometrics* 167, 330-344.

[11] Fan, Y., Sherman, R., Shum, M., 2014. Identifying Treatment Effects under Data Combination. *Econometrica* 82(2), 811-822.

[12] Fan, Y., Wu, J., 2010. Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and its Confidence Sets. *Review of Economic Studies* 77, 1002-1041.

[13] Firpo, S., Pinto, C., 2015. Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures, forthcoming in *Journal of Applied Econometrics*.

[14] Firpo, S., Ridder, G., 2008. Bounds on Functionals of the Distribution of Treatment Effects. IEPR Working Paper No. 08-09.

[15] Frank, M. J., Nelsen, R. B., Schweizer, B., 1987. Best-Possible Bounds on the Distribution of a Sum—a Problem of Kolmogorov. *Probability Theory and Related Fields* 74, 199-211.

[16] Guerre, E., Sabbah, C., 2012. Uniform Bias Study and Bahadur Representation for Local Polynomial Estimators of the Conditional Quantile Function. *Econometric Theory* 28(1), 87-129.

[17] Hahn, J., 1998. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica* 66, 315-331.

[18] Hahn, J., 2004. Functional Restriction and Efficiency in Causal Inference. *Review of Economics and Statistics* 86, 73-76.

[19] Hall, P., van Keilegom, I., 2009. Nonparametric "Regression" when Errors are Positioned at End-points. *Bernoulli* 15, 614–633.

[20] Hall, P., Sheather, S.J., 1988. On the Distribution of a Studentized Quantile. *Journal of the Royal Statistical Society, Series B* 50, 381–391.

[21] Hardy, G. H., Littlewood, J. E., Polya, G., 1934. *Inequalities*, Cambridge University Press.

[22] Heckman, J., 1990. Varieties of Selection Bias. *American Economic Review* 80(2), 313-318.

[23] Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998a. Characterizing Selection Bias Using Experimental Data. *Econometrica* 66, 1017-1098.

[24] Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998b. Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* 65, 261–294.

[25] Heckman, J., Smith, J., Clements, N., 1997. Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts. *Review of Economic Studies* 64, 487-535.

[26] Heckman, J., Vytlacil, E., 1999. Local Instrumental Variable and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences* 96, 4730-4734.

[27] Heckman, J., Vytlacil, E., 2001. Local Instrumental Variables, In: Hsiao, C., Morimune, K., Powells, J. (Eds.), Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya. Cambridge University Press, Cambridge, pp. 1-46.

[28] Heckman, J., Vytlacil, E., 2005. Structural Equations, Treatment, Effects and Econometric Policy Evaluation. *Econometrica* 73(3), 669-738.

[29] Heckman, J., Vytlacil, E., 2007a. Econometric Evaluation of Social Programs. Part I: Causal Models, Structural Models and Econometric Policy Evaluation, in *Handbook of Econometrics* 6B, 4779-4874.

[30] Heckman, J., Vytlacil, E., 2007b. Econometric Evaluation of Social Programs. Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments, in *Handbook of Econometrics* 6B, 4875-5143.

[31] Hirano, K., Imbens, G. W., Ridder, G., 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71, 1161-1189.

[32] Horowitz, J. L., Manski, C. F., 2000. Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association* 95, 77-84.

[33] Imbens, G. W., Manski, C. F., 2004. Confidence Intervals For Partially Identified Parameters. *Econometrica* 72, 1845–1857.

[34] Kong, E., Linton, O.L., Xia, Y., 2010. Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory* 26, 1529–1564.

[35] Lee, M. J., 2005. *Micro-Econometrics for Policy, Program, and Treatment Effects*, Oxford University Press.

[36] Makarov, G. D., 1981. Estimates for the Distribution Function of a Sum of two Random Variables When the Marginal Distributions are Fixed. *Theory of Probability and its Applications* 26, 803-806.

[37] Manski, C. F., 1997. Monotone Treatment Effect. *Econometrica* 65, 1311-1334.

[38] Moynihan, R., Schweizer, B., Sklar, A., 1978. Inequalities among Binary Operations on Probability Distribution Functions, In: Beckenbach, E.F. (ed.) General Inequalities, vol. 1, pp.133–149, Birkhauser Verlag, Basel.

[39] Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric Estimation of Index Coefficients. *Econometrica* 55, 1403–1430.

[40] Rachev, S. T., Rüschendorf, L., 1998. *Mass Transportation Problems*, Vol. I, Berlin Heidelberg New York, Springer.

[41] Rosenbaum, P. R., Rubin, D. B., 1983a. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society, Series B* 45, 212-218.

[42] Rosenbaum, P. R., Rubin, D. B., 1983b. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1), 41-55.

[43] Rüschendorf, L., 1982. Random Variables with Maximum Sums. *Advances in Applied Probability* 14(3), 623-632.

[44] Stoye, J., 2009. More on Confidence Intervals for Partially Identified Parameters. *Econometrica* 77, 1299-1315.

[45] Stoye, J., 2010. Partial Identification of Spread Parameters. *Quantitative Economics* 1, 323-357.

[46] Tchen, A. H., 1980. Inequalities for Distributions with Given Marginals. *Annals of Probability* 8, 814-827.

[47] Vijverberg, W. P. M., 1993. Measuring the Unidentified Parameter of the Extended Roy Model of Selectivity. *Journal of Econometrics* 57, 69-89.

[48] Williamson, R. C., Downs, T., 1990. Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds. *International Journal of Approximate Reasoning* 4, 89-158.